

PERIODICAL STACKS

JUL 1 1996

JUN 28 1996

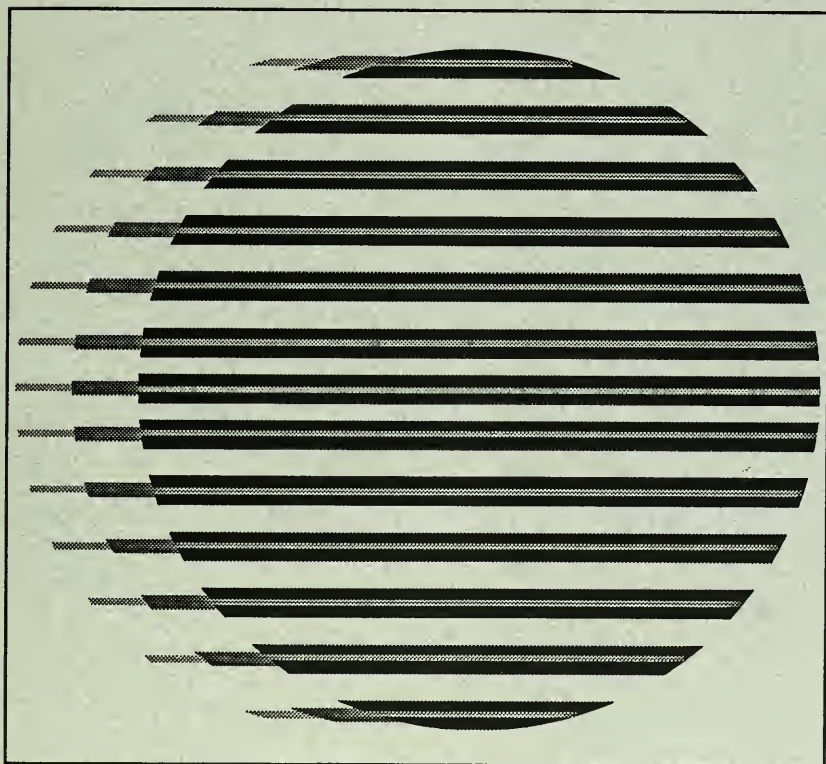
IASSIST

Q U A R T E R L Y

VOLUME 19

Fall 1995

NUMBER 3



Printed in the U.S.A.

IASSIST QUARTERLY



The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors

The QUARTERLY is published four times per year. Articles and other information should be typewritten and double-spaced. Each page of the manuscript should be numbered. The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". *Journal of the American Society for Information Science* 30(2):77-82, March 1979. If the contribution is an announcement of a conference, training session, or the like, the text should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event. Book notices and reviews should not exceed two double-spaced pages. Deadlines for submitting articles are six weeks before publication. Manuscripts should be sent in duplicate to the Editor: Laura Bartolo, Libraries & Media Services, Kent State University, Kent, Ohio 44242. (216) 672-3024. Email: LBARTOLO@KENTVM.KENT.EDU. Book reviews should be submitted in duplicate to the Book Review Editor: Daniel Tsang, Main Library, University of California P.O. Box 19557, Irvine, California 92713 USA. (714) 856-4978 E-Mail: DTSANG@ORION.CF.UCLD

Title: Newsletter - International Association for
Social Science Information Service and
Technology
ISSN - United States: 0739-1137 Copyright 1985 by
IASSIST. All rights reserved.

CONTENTS

Volume 19

Number 3

Fall 1995

FEATURES

- 4** The Development of Software to Facilitate
Use of Archived Data Sets
by Josefina J. Card and Elizabeth A. McKean
- 12** INCORE Metadatabase Server - Technical
Aspects and Constraints
by Patrick Curran
- 20** Locating and Accessing Data and
Information on the Internet: Methods and
Organizational Impacts
by Christopher Davis
- 28** SOSIG - a move towards subject-based
services.
by Nicky Ferguson
- 31** WWW: What do Researchers Want?
by Jim Henderson
- 36** Georeferenced Population Data
by Hendrik Meij and Robert Chen

The Development of Software to Facilitate Use of Archived Data Sets

by Josefina J. Card and Elizabeth A. McKean'
Sociometrics Corporation

A variety of forces have converged to promote research based on secondary analysis of existing social science data sets. A growing number of federal agencies have officially encouraged data sharing by requesting, often requiring, that their grantees place data sets collected with public funds in the public domain. At the same time, declines in university and federal research budgets have put expensive primary data collection out of the reach of many social scientists. Advances in microcomputer technology have allowed powerful data analyses to be performed quickly and economically. Data archive centers dedicated to the preparation of data sets for public use have also begun to emerge.

Several challenges await the data archivist working to provide users with clean, useable data for secondary analysis. The user must be provided with data of high quality, documented in clear and comprehensive fashion. While paper documentation retains its value, paperless (electronic) documentation is becoming increasingly important, in light of burgeoning use of the Internet and mass storage media such as the CD-ROM. Hand-in-hand with the growth of the national movement toward secondary data analysis comes the need for powerful yet user-friendly ways to search through the massive amounts of available data and documentation to retrieve studies or variables of interest. To reduce hard-disk storage burdens and statistical analysis time, such search and retrieval of variables would ideally be linked with data extract capabilities, so that analysis files containing only variables or cases of interest to the user can be created on demand.

This article documents the latest advancements of one data archive center, Sociometrics Corporation, in meeting these challenges. The Sociometrics Data Library currently houses five topically-focused data archives: the Data Archive on Adolescent Pregnancy and Pregnancy Prevention, the American Family Data Archive, the Data Archive of Social Research on Aging, the Maternal Drug Abuse Data Archive, and the AIDS/STD Data Archive. Together, these five data archives include over 200 data sets, which have been chosen for technical quality, scientific merit, substantive utility, relevance to social policy, demand for secondary data analysis, and disciplinary balance by a panel of experts in each archive's substantive field. Each data set in each topically-focused archive is made publicly available with a printed and bound user's guide and a standard set of machine-readable files—raw data, SPSS and SAS program statements that fully document the variables and values in the data file, an SPSS dictionary, and SPSS frequencies—with the explicit goal of providing the user with clear documentation and ready-to-use data files (see Card and McKean, 1993 for a discussion of standard file preparation).

This paper will focus on the recent development of three software features accompanying the data sets which simplify secondary analysis for social scientists. While the examples used will be drawn from the AIDS/STD Data Archive, the software described is generic to the data archives comprising the Sociometrics Data Library.

Software to Facilitate the Selection and Analysis of Variables

Search and Retrieval Software. The first feature, Search and Retrieval software, allows the user to examine the contents of an entire topically-focused data archive and retrieve variables using a variety of search strategies: (1) searches by full-text keyword, including variable names, variable labels (question descriptors), and value labels (response descriptors); (2) searches by substantive Topic and Type codes that have been assigned to each variable during the archiving process; and (3) searches by study name, author, or assigned data set number. Standard Boolean operators such as "and," "or," and "not" can be used to conduct any search.

Electronic Instrument-Variable Link. The second feature, an electronic link between study variables and graphic images of the data collection questionnaire, allows the analyst to select variables and view the instrument pages associated with the selected variables. Alternatively, the user may browse through the entire collection of instruments page-by-page, or search the instrument database by substantive keywords. The instrument-variable link allows analysts to examine questionnaire skip patterns and item context on-screen, a process which enhances the variable selection process and reduces the need for paper copies of instruments.

Data Extracting. The third feature, Data Extract software, allows the user to produce, with a few keystrokes, SPSS or SAS program statements for any subset of selected variables and then create an active or system file for statistical analysis. This

capability permits analysis of even the largest of data sets to be conducted on most microcomputers. It also saves significant preparation time in writing and re-writing SPSS and SAS program statements to define variables used in a given analysis.

In the following section, the capabilities of these Search & Retrieval, Instrument Link, and Data Extract software programs will be illustrated with a sequence of searches from the AIDS/STD Data and Instrument Archive, which contains over 14,000 variables from 11 major investigations of the incidence and prevalence of specific sexual behaviors, contraceptive use and STD preventive behavior, AIDS/STD knowledge, and attitudes regarding contraception and STD prophylaxis.

Search and Retrieval of Variables from the AIDS/STD Data and Instrument Archive

In the first example, a search by topic across the 11 studies comprising the AIDS/STD Archive illustrates how the Search and Retrieval software can be used to find all variables indexed under the substantive Topic of hiv/aids. After initiating the program and pressing the F3 select menu to request a search by topic, the F4 search menu is used to perform the search using the wildcard query topic? (Figure 1).

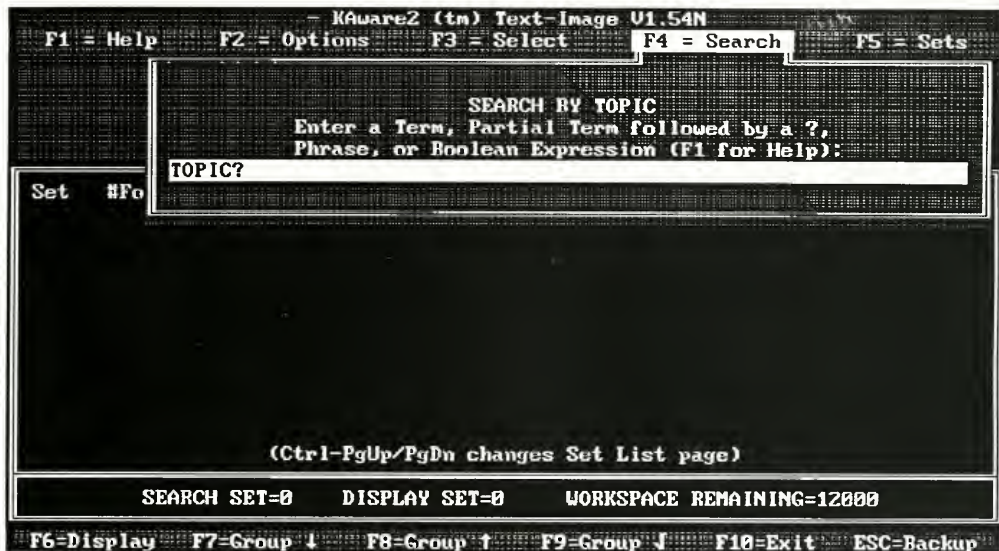


Figure 1

Wildcard searches such as the topic search requested in Figure 1 produce a scrollable menu of retrieved items. In this example, the 62 substantive Topics available in the AIDS/STD archive are displayed in the menu. The third of the six screens comprising this menu is shown in Figure 2.

Figure 2 shows that 630 variables in the 11 studies comprising the AIDS/STD archive have been indexed under the topic of hiv/aids. Descriptions of each of these "hit" variables can be viewed by highlighting the hiv/aids topic line in the scrollable menu and pressing enter. Variables records for all 630 "hit" variables will be retrieved, and the first variable record in the set will be displayed on the screen. Figure 3 shows the first variable record from the topic = hiv/aids search set.

As Figure 3 shows, the variable text record provides the variable name and label (Line 1), study name (Line 3), author or investigator names (Line 4), topic and type codes assigned to the variable (Lines 5 and 6), and value labels (Line 7 ff.). The text record also contains the following on-screen instructions (Line 2) for viewing the instrument page containing the original item from which the variable was derived: "press alt + i to see image of actual questionnaire." In Figure 4, the instrument page for hvb01021: a:7a been to hiv testing site is shown in the upper left corner of the screen. The lower right portion of the Figure 4 screen contains instructions for zoom enlargement of the instrument page image, which can be



Figure 2

magnified to one of three sizes and printed out. In Figure 5 the instrument page for variable hvb01021 is magnified at zoom level 2 and has been cropped to fit the page.

To allow the user to quickly locate the original item on the graphic page image, all variable labels include the questionnaire item number. In this example, variable hvb01021: a:7a been to hiv testing site, is from item 7.a. on the "A" questionnaire from Data Set 01, *The California Survey of AIDS Knowledge, Attitudes and Behavior: 1987*. Figure 5 shows that item 7.a. is



Figure 3

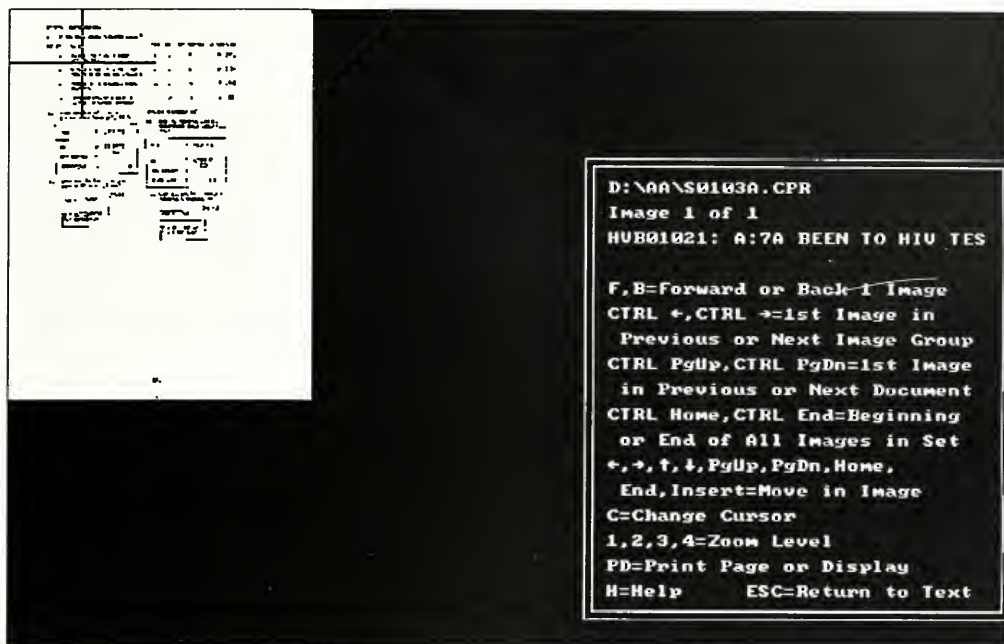


Figure 4

part of a compound question assessing AIDS-related behaviors.

Because one of the primary goals of a search and retrieval session is to select a subset of variables for analysis, the next step in our example will show how a search set of demographic variables including age, race, sex, and community of residence can be combined with the set of 630 hiv/aids variables already retrieved, in order to produce a prototypic set of analysis variables. Under the **f4 search by topic** function, the user can select and retrieve multiple topics at one time with the **f9 (group +)** key. Our second topic search retrieves all variables in the AIDS/STD Archive that assess age, race or ethnicity, gender or gender role, neighborhood or community, and region or state — a total of 717 variables indexed under these five different topics. Figure 6 shows the fifth of a set of six topic-menu screens, showing how such selection was done for two of the five topics (**race/ethnicity and region/state**).

The 717 "hit" variables from this second search covering five demographic topics are combined with 630 hit variables from the first HIV/AIDS-topic search using the Boolean operator **or** (Figure 7).

A new set of 1,347 variables results (Figure 8, "Set 3"). To save the selection of demographic and HIV/AIDS variables comprising Set 3 in a file that can be used by Sociometrics' Data Extract software, the user simply selects the option "Transport a Set" from the **f5 sets** menu and provides a filename for the transported set (Figure 8).

Transforming a Variable Search Set into a Statistical Analysis Package Command File

The Data Extract software uses the transported set file to create extract command files in user's choice of the SPSS or SAS statistical analysis package. Figure 9 shows the on-screen summary produced after the sample transport file has been read. Each study from the AIDS/STD Data and Instrument Archive contains some of the demographic and HIV/AIDS variables from the search.

(ASK ALL RESPONDENTS:)					
7. In the past year, have you . . . ?					
(ROTATE SERIES)		YES	NO	NOT SURE/DC	REFUSED/NA
a.	BEEN TO AN AIDS ANTIBODY TESTING SITE	1	2	8	9 (32)
b.	RECEIVED CARE AT A MILITARY SCREENING OR MEDICAL FACILITY	1	2	8	9 (33)
c.	DONE TIME IN A CORRECTIONAL FACILITY	1	2	8	9 (34)
d.	HIGH TREATED AT A DRUG OR ALCOHOL TREATMENT PROGRAM	1	2	8	9 (35)

Figure 5

Extract command files may be produced for each data set in turn. The program requires between 30 seconds and 3 minutes to produce an ASCII command file that will create an SPSS/PC+, SPSSx, or SAS active file with variable names, variable labels, and value labels.

Figure 10 shows a sample extract SPSS-PC+ command file for Data Set 01, the *California Survey of AIDS Knowledge, Attitudes, and Behavior: 1987*. To save space, only a sample of variables from this file is presented; lines edited out of the

SPSS-PC+ command file in Figure 10 are noted with a series of dots (....). The resulting extract command file can be used to

```

- KAware2 (tm) Text-Image V1.54N
F1 = Help      F2 = Options    F3 = Select    F4 = Search    F5 = Sets

SEARCH BY TOPIC
TOPIC: PF PSYCHOLOGICAL FUNCTIONING/DEVELOPMENT [99]
TOPIC: PP INTER-PARTNER RELATIONSHIPS [392]
TOPIC: PR PROSTITUTION [443]
TOPIC: RA RACE/ETHNICITY [173]
TOPIC: RC RECREATION [26]
TOPIC: RE REGION/STATE [51]
TOPIC: RL RELIGION [88]
TOPIC: RS RECEIPT OF HEALTH/MENTAL HEALTH/SOCIAL SERVICES [49]
TOPIC: SA SUBSTANCE USE [673]
TOPIC: SD SEXUALLY TRANSMITTED DISEASE, OTHER [494]
(PAGE 5 OF 6)

(Ctrl-PgUp/PgDn changes Set List page)

SEARCH SET=0    DISPLAY SET=0    WORKSPACE REMAINING=12000

F6=Display  F7=Group ↓  F8=Group ↑  F9=Group ↓  F10=Exit  ESC=Backup

```

Figure 6

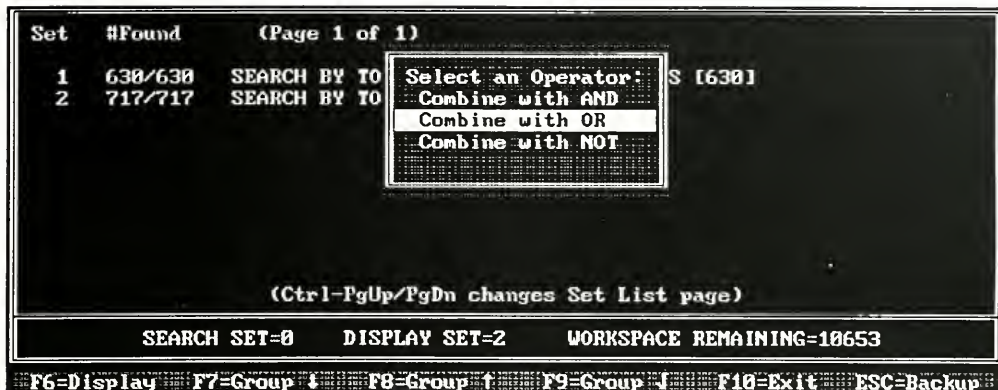


Figure 7

create an SPSS/PC+ active file, or with minor editing, a system file on which analyses of the 100 Age, Race, Gender/Gender Role, Neighborhood/Community, Region/State, and HIV/AIDS variables in the *California Survey of AIDS Knowledge, Attitudes, and Behavior: 1987* can be performed with ease.

Next Step

The last decade has witnessed paradigmatic changes in the way social science data sets are stored, delivered to users, and analyzed. Many of these changes have been brought about by rapid technological developments in microcomputer hardware and software, optical storage devices, and the growth of the Internet and on-line digital libraries. This paper has briefly described one state-of-the-art data archive consisting of high quality data in a field of current interest. The AIDS/STD Data and Instrument Archive is representative of all data archives produced at Sociometrics, which combine high quality,

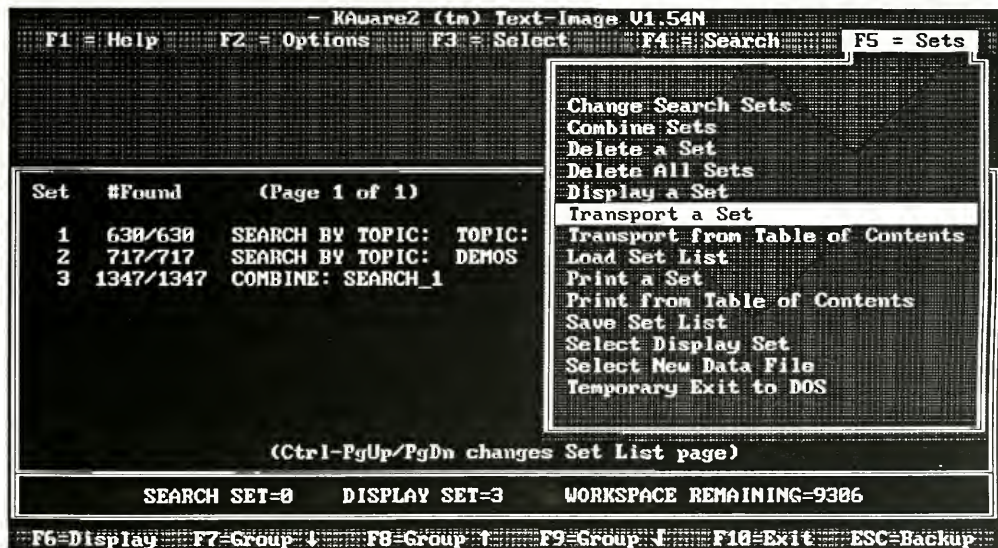
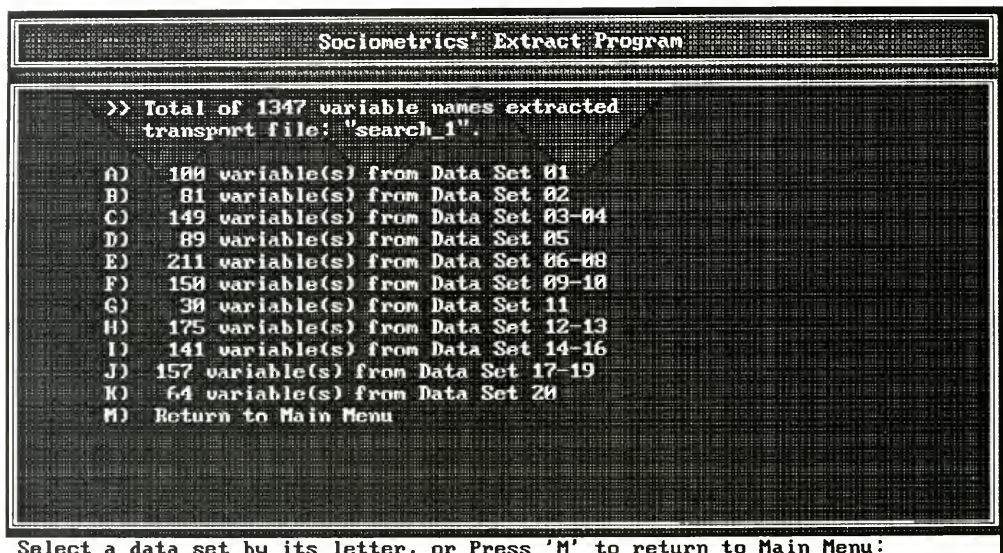


Figure 8

paperless documentation with powerful, yet easy-to-use software for variable search and retrieval, viewing of the original questionnaire page containing the item, and data extraction.

As the need for data sets for secondary analysis grows, social scientists will demand more sophisticated methods of data storage and retrieval. Data archivists will continue to face the challenge of providing users with increasingly well-prepared, paperless, and machine-searchable data collections. In anticipation of user needs, Sociometrics is developing several new enhancements for future data archives. First, programming is being developed for electronic links between variables and descriptive statistics as well as technical notes such as skip logic, scale variables, case weights, and other study-specific information associated with each variable in an archive. Because of the overwhelming amount of information that typically accompanies a data set, it is important that accurate and complete documentation at the level of the individual variable also be accompanied by print-on-demand capabilities. Providing print-on-demand capabilities for user-selected portions of documentation (e.g., user guide sections, questionnaire pages) in a variety of character formats including ASCII, Microsoft Word, WordPerfect and Postscript, constitutes a second forthcoming technological enhancement. Finally, Sociometrics is preparing

Figure 9



to launch SOCINET, an on-line Data Library service, available 24 hours a day to Internet users.

It is hoped that these developments will move the field of data sharing and secondary data analysis ever closer to the vision of the successful enterprise-of-the-future described by Stanley M. Davis in his book *Future Perfect*: the ability to meet users' needs—in this case their data information needs—any time, any place, any where (Davis, 1987).

References

Card, J. J. and McKean, E. A. (1993). Harnessing advances in technology: New opportunities for research and teaching in social gerontology. *Gerontology & Geriatrics Education*, 14,2, 63-76.

Davis, Stanley M. (1987). *Future Perfect*. Reading, Massachusetts: Addison-Wesley.

1. We wish to thank our colleagues, Drs. Kathryn Muller, Bill Farrell, and Eric Lang, for their comments on an earlier draft of this paper.

* Sociometrics Corporation (Los Altos, CA)
* AIDS/STD Data Archive, Data Set 01
* SPSS/PC+ Program Statements for
* The California Survey of AIDS Knowledge,
* Attitudes, and Behavior: 1987.

DATA LIST FILE = 'std01.raw'

/

GRS01003 20-27

/

/

HVB01021 20-27

/

HVC01031 28-35

HVC01032 36-43

HVC01033 44-51

HVC01034 52-59

HVC01035 60-67

HVC01036 68-75

...

VARIABLE LABELS

GRS01003 "** Respondent sex"

HVB01021 "A:7 a Been to HIV testing site"

HVC01031 "A:12 a AIDS Risk: Use public facilities"

HVC01032 "A:12 b AIDS Risk: Go to school w/PWA child"

HVC01033 "A:12 c AIDS Risk: Donate blood"

HVC01034 "A:12 d AIDS Risk: Live near hospital w/AIDS patients"

....

VALUE LABELS

GRS01003

1 "MALE"

2 "FEMALE"

HVB01021

1 "YES"

2 "NO"

8 "NOT SURE/DONT KNOW"

9 "REFUSED"

....

FINISH.

INCORE Metadatabase Server - Technical Aspects and Constraints

by Patrick Curran¹

INCORE, University of Ulster, N. Ireland.

The INCORE server was set up by the University of Ulster and the United Nations University to act as a central resource for academics, policy-makers and others concerned with conflict resolution and ethnicity across the globe. The server utilises Internet resources to make as widely available as possible information on ethnic conflict. However, like a lot of tools the Internet can be used for more than one purpose. In one sense it is a very flexible and open resource that allows users to log into numerous systems around the world, transferring files, searching file systems and executing programs. However once you are connected to the Internet it is also a way for other people to get into your system. If you restrict your workstation's access only to parts of the Internet, then you may find that there is a lot of useful resources and facilities that you can no longer access and use. In order to take advantage of the Internet you must be a part of it. However, as this paper will try to demonstrate, by doing this you put your computer at risk, so you need to constantly keep abreast of the security holes in the network software installed in your system and protect it.

Background

INCORE [2] was established by the University of Ulster (UU) and the United Nations University (UNU) to provide a systematic approach to the study of ethnic conflict and to encourage links between research, training, policy, practice and theory.

The INCORE metadatabase server [3] was set up to act as a central resource for academics, policy-makers and others concerned with conflict resolution and ethnicity across the globe, but particularly to those operating in areas of conflict. It is central to INCORE's mandate that it serve those who have most difficulty in getting access to information. In this respect the Internet can be both the most and least useful medium for connecting people to information. The advantage of placing information on the Internet is that it is almost instantaneously available, uncensored, to anyone anywhere in the globe who is connected to the net. The disadvantage is that, unless one is connected to the Internet, which involves expensive hardware, technical expertise and effective maintenance, this information is inaccessible. INCORE has to be concerned not merely to provide a state of the art server but also to ensure that as much of the information as possible on the server is accessible to those with the least connectivity to the Internet, through E-mail and FTP as well as through gopher and World Wide Web. However in setting up this server we need to be aware of adopting a responsible attitude towards security in order to maintain a continued uninterrupted service. We need to consider where our system is at risk [4], where the possible threats and breaches of security are coming from and how to prevent these becoming problems.

Introduction

As the volume of Internet use increases dramatically (30,000 plus interconnected networks with 2.5 million or more connected computers daily swap gigabytes of information based on nothing more than a digital handshake with a stranger) [5, 6] and as we connect our organisation's networks to thousands of other computer networks, serious security issues arise. There are a lot of considerations for the system administrator and the following list outlines some of the main issues :

- Unauthorised users accessing data and information on our system
- Authorised users causing inadvertent or malicious damage
- Interception of data transmission
- Virus attacks
- Backup and storage policy

It is difficult to fully understand and appreciate just how much the Internet depends on collegial trust and a group effort by the whole community. One technique, for instance, that intruders use is to break into a chain of computers, e.g., break into A, use A to break into B, B to break into C, etc., thus covering their tracks. So, it would be very unwise and foolish to think that your little part of the network is safe because you believe that there is very little information there of a nature that would entice someone to break into it. Even if there is nothing of use on your computer, it could prove a worthwhile intermediate staging post for someone who wants to break into other more useful systems.

In the early days of the ARPAnet only researchers had access to the net, and they shared a common set of goals and ethics. Data packets were forwarded along network links from computer to computer. A packet may have made a number of hops and every intermediate machine could read its contents. Nowadays many Internet packets start their journey on a local area network (LAN), where privacy is even less protected. Only a gentleman's agreement assures the sender that the recipient and no-one else will read the message. The lack of security on the ARPAnet did not bother anyone, because that was part of the package. As the Internet developed and expanded, the user population began to change, with a lot of the newcomers having little idea of the importance of the complex social contract guiding the use of this new and exciting tool. Nowadays anyone with a computer, a modem and a small monthly connection fee can have a direct link to the Internet and be subject to break-ins or launch attacks on others.

Every day computer networks and hosts are being broken into with varying levels of sophistication. While it's generally believed that most break-ins succeed due to weak passwords, there are advanced and sophisticated techniques that are more difficult to detect. The article looks firstly at the shortcomings of the password mechanism (concentrating on the Unix system) and then discusses the more sophisticated techniques available to intruders, including some examples of the misuse of commonly used Internet resources, e.g., ftp, sendmail, telnet, rlogin, rsh, etc. It also analyses the range of responses to network intrusion techniques, from software policing solutions like Kerberos and COPS to the hardware solution of firewall installation.

Security Breaches

System administrators can safely configure workstations on their network to allow connections to other workstations. They can also set up their network file system to export widely used file directories to "world", allowing everyone to read them. It doesn't take much imagination to see what can happen when such a trustworthy environment opens up its digital doors to the Internet. Suddenly, 'world' means the entire globe and "any computer on the network" means "every computer on any network". There have been a lot of computer security problems and breaches of security in recent years, some more serious than others. Some of the more widely known incidents include break-ins on NASA's SPAN network [7] and the IBM "Christmas Virus", but the most widespread breach of network security occurred in 1988 when the Internet came under attack from within, later to be known as the Internet Worm Incident.

Internet Worm

On November 2, 1988, a self-replicating program, called a worm* appeared on the Internet [8]. This program copied itself from machine to machine, causing the machines it infected to labour under huge loads, and denying service to the users of those machines. The program spread quickly, and while many system administrators were aware that something like this could theoretically happen (the security holes exploited by the worm were well known) the scope of the worm's break-ins came as a great surprise to many.

The worm itself did not destroy any files, steal any information (other than account passwords), intercept private mail, or plant other destructive software [9]. However, it did manage to severely disrupt the operation of the network. Several sites, including parts of MIT, NASA's Ames Research Centre and Goddard Space Flight Centre, the Jet Propulsion Laboratory, and the Army Ballistic Research Laboratory, disconnected themselves from the Internet to avoid recontamination. In addition, the Defense Communications Agency ordered the connections between the MILNET and ARPANET shut down, and kept them down for nearly 24 hours [10]. Ironically, this was perhaps the worst thing to do, since the first fixes to combat the worm were distributed via the network.

This incident was perhaps the most widely described computer security problem ever. The worm was covered in many newspapers and magazines around the country including the New York Times and most computer oriented technical publications.

Security Considerations

The incidents above demonstrate quite clearly that computer security is an important topic. Every day computer networks and hosts are being broken into with varying levels of sophistication. When you hear the term "security" the first thing that comes to mind is the password. However, while it's generally believed that most break-ins succeed due to weak passwords, there are, moreover, a large number of unauthorised attacks that use more advanced and sophisticated techniques. Less is known about these latter techniques because they are more difficult to detect. We will look firstly at the advantages and disadvantages of the password mechanism (concentrating on the Unix system [11]) and then discuss the more sophisticated techniques available to intruders.

Passwords

An underlying goal of the Unix system has been to provide password security at a minimal inconvenience to the users of the system. For example, those who want to run a completely open system without passwords, or to have passwords only at the option of the users, can do so, whilst those who require all their users to have passwords gain a high degree of security against penetration of their system by unauthorised users. A good password system must be able not only to prevent access to the system by unauthorised users, but it must also prevent users already logged on from doing things that they are not authorised to do. The "super user" password, e.g., is especially critical because it allows all sorts of permissions and provides unlimited access to all system resources.

Passwords are important because they are generally the first line of defence against interactive attacks. In simple terms, if a cracker[®] cannot interact with your system, and he has no access to read or write to the information contained in the password file, then he has almost no avenue of attack left open to break your system. Unfortunately the Unix passwd program [12] doesn't place a lot of restrictions on what might be used as a password. It generally requires 5 lowercase letters (or 4 characters) but if the user insists on using a shorter password (by entering it three times) the program allows it.

The object when choosing a password is to make it difficult for a cracker to make educated guesses, thus leaving him with no alternative but to try every possible combination of letters, characters, special characters and numbers. A search of this sort, even on the most powerful computer would take at least 100 years to complete.

Robert Morris and Ken Thomson carried out an interesting survey determining typical users' habits in the choice of passwords [13]. Out of 4,000 passwords 16% contained 3 characters or less, and 86% were what could generally be described as insecure. Grampp & Morris [14] in another experiment showed that by trying the 20 most common female names, followed by a single digit, at least 1 password was valid in each of the machines surveyed. They also found that by trying variations of the login name, user's first and last name, and a list of nearly 1800 common first names, that up to 50% of the passwords on any given system can be cracked in a matter of 2 or 3 days.

There are ways to improve the security of using passwords. According to Schweitzer [15] a high quality password has the following characteristics :

- At least eight characters length
- The password is randomly generated
- The password has no personal relationship to the user, his family or job
- The password must be kept secret
- It must be changed at least every three months

The Internet worm, in trying to break into new systems attempted to crack user passwords [7]. First of all it tried simple choices (user names, names, etc) and then tried an internal dictionary of 432 words. If all failed it tried going through the system dictionary /usr/dict/words, trying each in turn. So, by adhering to the guidelines above, you will make it very difficult for an intruder to break into your system using the first line of defence.

Other problems associated with passwords are :

Expired accounts : Accounts lying around due to people leaving the organisation. These cause problems because since nobody is using the account anymore it's unlikely that a break-in will be noticed. A simple preventative measure is to place an expiration date on every account. If there is any doubt about an account, then replacing the encrypted password with an asterisk (*) will make it impossible for anyone to log into the account.

Guest accounts : These are usually made available for expediency. They should only be used for the period required and deleted immediately. It's important not to give them simple passwords, e.g., "guest" or "temp".

Network Security

Besides cracking passwords on machines crackers have other means at their disposal, the success of which depends on your awareness and adoption of preventative measures. It's important to be careful about techniques that bypass password requirements. There are two common ones, i.e., the .rhosts and the hosts.equiv files :

hosts.equiv : the file /etc/hosts.equiv can be used to indicate trusted hosts. If a user remotely logs into your system using rlogin and does so from a host listed in this file, access is permitted without requiring a password. The default

file has the entry '+' in a single line indicating that every host should be considered a trusted host. This could prove a major security problem since hosts outside the local organisation should never be trusted. Only specific host names should be included in the file, or the file deleted altogether.

.rhosts: allows access to specific host-user combinations. Each user may create a *.rhosts* file in his/her directory, allowing access to their account without supplying a password. For example, the entry - host.com fred - tells the computer on which the file resides to bypass password requirements when it sees someone logging in from the login name "fred" on the machine "host.com". This means, obviously, that anyone who manages to break into fred's account can also break into this machine.

The Internet worm made use of the trusted host concept to spread itself throughout the network [8].

Secure Terminals

Unix introduced the concept of a "secure" terminal, which prohibits 'root' from logging in from a non-secure terminal. The file */etc/ttyab* controls which terminals are considered secure. The default is to consider all terminals secure allowing root to login from anywhere in the network. A more secure configuration would be to consider as secure only directly connected terminals, or only the console device. The most secure method is to remove the secure designation from all terminals including the console, so users with 'root' privilege must first login as themselves and then use the 'su' command.

NFS

The Network File System (NFS) allows several hosts to share files over the network, generally used to provide file server access to diskless workstations in a small network. The file */etc/exports* lists which file systems are exported. This file contains access specifications, e.g. :

<i>root=keyword</i>	specifies the list of hosts allowed "super-user" access to the files in the named file system
<i>access=keyword</i>	specifies the list of hosts that are allowed to mount the named file system.

For example, the line - /export/root/client -access=client,root=client - allows the host "client" to access the named file system with root privileges. If the file isn't properly configured then anyone on the Internet may have access to your files via NFS, whether you trust them or not.

E-Mail

Whilst this is one of the net's basic services, forging e-mail is a trivial exercise. An electronic letter consists simply of a text file with a header specifying the sender, receiver, subject, date and routing information, followed by a blank line and the message body. Mail programs fill in the header lines with routing information, but there's nothing to stop a malicious person inserting whatever he wants into the mail. Protocols have been developed to verify the source of e-mail messages, but spoofer's are also improving their techniques. Some systems bar connection to their system from untrustworthy parts of the Internet. The problem with this strategy is that the trusted "domain name servers" they rely on are just ordinary computers, and as such are also vulnerable to deception or intrusion. A cracker can modify the name server's database so that it tells any computer querying it that the address belonging to e.g., "cracker@breach.com" is instead that of "president@whitehouse.gov". A computer allowing connection from whitehouse.gov will allow the cracker in as well.

The "sendmail" bug has reappeared time and time again over the years, due to the fact that most mail programs make it possible to route messages not only to users but also directly to particular files or programs. People, e.g., forward mail to a program called "vacation" which sends a reply telling the correspondent that the recipient is out of town. Other people route mail through filter programs that forward the message to various locations. This same mechanism can thus be subverted to send e-mail to programs that are designed to execute shell scripts. Such a script could cause a copy of the receiving computer's password file to be sent to an intruder for analysis, or simply wreak havoc on the recipient's file system.

Intrusion Techniques

Many system administrators are often unaware of the dangers presented by anything beyond the most trivial attacks. The purpose of this section is to present a few of the techniques available to the system cracker in his efforts to gain access to a shell process on a Unix host [16]. There are so many methods and techniques that it would be impossible to cover them all in this paper. However, I will try to outline some of the most common techniques. Fig. 1 illustrates some of the more useful services and facilities that an intruder may avail of to ferret out information about a system and set up an interactive link with that system.

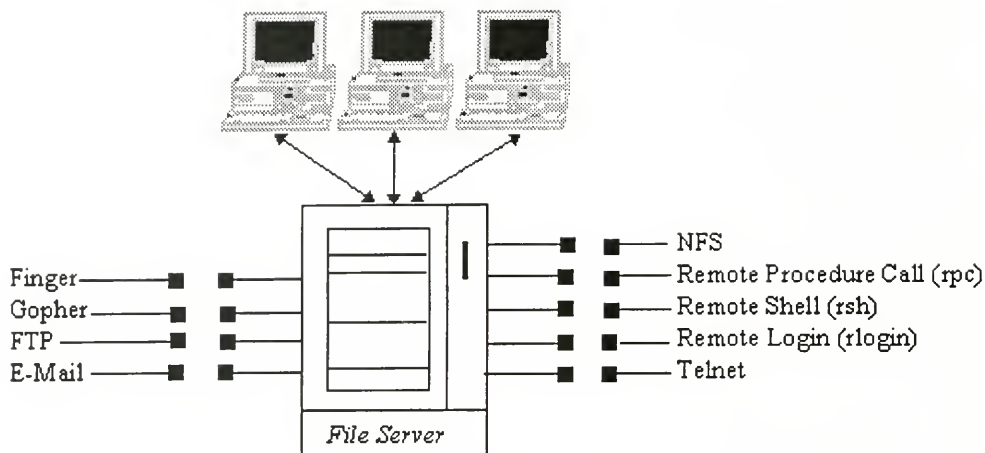


Fig. 1 : Network Intrusion Techniques

The finger services, provided by the finger program outputs information about the users logged on to the system, and the fingerd program extends this facility to remote hosts, e.g.,

```
host% finger@incore.ac.uk
Login  Name  TTY  Idle When  Where
pat    P.Curran  co   21  Fri 10:38 i  inf.ac.uk
```

The most revealing information divulged are the account names, home directories and the host they last logged in from. This information can be supplemented by using the rusers command (with the -l flag), which produces, e.g. :

Login	Home-dir	Shell	Last login, from where
root	/	/bin/sh	Fri Apr 15 from inf.ac.uk
pat	/home/pat	/bin/csh	On since Thur Apr 14 from inf.ac.uk
guest	/export/guest	/bin/sh	Never logged in
ftp	/home/ftp		Never logged in

Finger is one of the most dangerous services because it is very useful for investigating and receiving information about possible target machines, especially when used in conjunction with other data. A bug in the fingerd program was exploited by the Internet worm to overrun the buffer that the daemon (background process that fingerd is intended to run as) used for input, thus altering the behaviour of the program.

The *showmount* command can be used on an NFS fileserver to display the names of all the hosts that currently have something mounted from the server. Running showmount on a target, e.g., reveals :

```
host% showmount -e incore.ac.uk
export list for incore.ac.uk
/export  (everyone)
/var     (everyone)
```

Since /export/guest is exported to the world, and this is the user guest's home directory, we have a possible break-in scenario. The intruder could mount the home directory of 'guest', and create a 'guest' account in his local password file. By putting a

.rhosts entry in the remote guest home directory he allows himself to login to the target machine without having to supply a password.

If the target machine has a '+' wildcard in its /etc/hosts.equiv file, then any non-root user with a login name on the target's password file can rlogin to the target without a password. So the intruder's next line of attack would be to login to the target host and modify the password file to allow root access :

```
host% rsh incore.ac.uk csh -i
host% ls -ldg /etc
drwxr-xr-x  8bin  staff  2048 Jul 24 18:02 /etc
host% cd /etc
host% mv passwd pw.old
host% (echo attack::0:1:instant root shell:::/bin/sh;cat pw.old)> passwd
host% ^D
host% rlogin incore.ac.uk -l attack
Welcome to the Incore Server !!
incore#
```

The rsh -i gets one on to the system but doesn't leave any traces in the wtmp or utmp system auditing files, making the remote shell invisible to the finger and who commands.

Going back to the finger output we can see that there is an 'ftp' account, which usually means that anonymous ftp is enabled. The File Transfer Protocol (ftp) allows users to connect to remote systems and transfer files back and forth. This can be an easy way to get access to a system as it is often misconfigured. In some cases the target machine has a complete copy of the /etc/passwd file in the anonymous ftp ~ftp/etc directory. If, however, this isn't the case then there are other avenues for the would-be attacker. If, for instance, the ftp directory is writable, an intruder can remotely execute a command, e.g., mailing the password file back to himself simply by creating a .forward file that executes a command when mail is sent to the ftp account.

If none of the methods above have worked then the attacker could use rpcinfo to see if the target is running NIS or NFS. Once you know the NIS domainname of a server you can get any of its NIS maps using a simple rpc query. Also, just like easily guessed passwords many systems use easily guessed NIS domainnames [17]. The showmount output usually divulges information on domainnames, which can then be checked with the 'ypwhich' command to see if the domainname exists, e.g.

```
host% ypwhich -d incore incore.ac.uk
Domain incore not bound
```

This proved unsuccessful, but if it was guessed properly it would have returned with the hostname of incore.ac.uk's NIS server. If an attacker has control of the NIS master then he effectively has control of the client hosts, and can execute arbitrary commands, e.g., mailing password files to himself.

Security Tools

There are a lot of tools available that deal with the security management aspects of computer systems, more than can be adequately covered in this article [18]. Because there are so many tools and techniques available to implement security controls you should, in the first place, identify the requirements of your network service and what you are willing to accept. The following sections provide only a flavour of the types of solutions available, from software based system auditing, authentication and checking techniques to the hardware solution of firewall installation.

Computer Oracle and Password System (COPS)

COPS [19] is a UNIX security status checker, written as a suite of shell scripts which forms an extensive security testing system. Basically it checks various files and software configurations to see if they have been compromised (e.g., edited to plant a trojan horse* or back door*), and checks to see that files have the appropriate modes and permissions set to maintain the integrity of your security level (making sure that your file permissions don't leave themselves wide open to attack). There's a rudimentary password cracker, and routines to check the file store for suspicious changes in setuid programs, and to identify software behaving in ways which could cause problems.

The current version of COPS makes a limited attempt to detect bugs that are posted in CERT advisories. Also, it has an option to generate a limited script that can correct various security problems that are discovered.

Kerberos

Kerberos [20, 21] is a DES-based encryption scheme that encrypts sensitive information, such as passwords, sent via the network from client software to the server daemon process. When a user logs in, Kerberos authenticates that user (using a password), and provides the user with a way to prove her identity to other servers and hosts scattered around the network. This authentication is then used by programs such as `rlogin` to allow the user to log in to other hosts without a password (in place of the `.rhosts` file). The authentication is also used by the mail system in order to guarantee that mail is delivered to the correct person, as well as to guarantee that the sender is who he claims to be. The overall effect of installing Kerberos and the numerous other programs that go with it is to virtually eliminate the ability of users to "spoof" the system into believing they are someone else.

Firewall

A firewall is a machine which is usually attached between your site and a wide area network. It provides controllable filtering of network traffic, allowing restricted access to certain internet port numbers (i.e., services that your machine would otherwise provide to the network as a whole) and blocks access to pretty well everything else. They are an effective "all-or-nothing" approach in dealing with external access security, and are fast becoming very popular, particularly with the rise in Internet connectivity [22].

The firewall doesn't send out routing information about the internal network, making the internal network "invisible" from the outside. It doesn't advertise routes which means that users on the internal network must log in to the firewall before accessing hosts on remote networks. Also, in order to remotely log in to a host on the internal network from the outside, a user must first log in to the firewall machine, which may prove to be inconvenient, but, nevertheless, more secure.

Outgoing e-mail is forwarded to the firewall machine before being delivered outside the internal network and the firewall receives all incoming e-mail, before redistributing it. It provides extra security by not mounting any file systems via NFS, or making any of its file systems available to be mounted. Password security is rigidly enforced and the firewall does not trust any other hosts regardless of where they are. Finally, anonymous ftp and other similar services is only provided by the firewall host, if at all.

The purpose of the firewall is to prevent crackers from accessing other hosts on your network which means that security must be strictly and rigidly enforced. It is important to remember that a firewall can't provide complete safeguards against intrusion - if someone manages to subvert the firewall then he can subsequently break into any host behind it.

Many organisations, and more recently universities, have adopted firewalls to examine the packets entering and leaving a domain and to restrict certain Internet connections. However, proposing a firewall and constructing it are two different things entirely. There are some things that you just can't do securely. Gopher and Mosaic, for instance, are two programs of a trusting nature that defy the attempts of a firewall design to provide safety. Additionally, a firewall must pass mail and

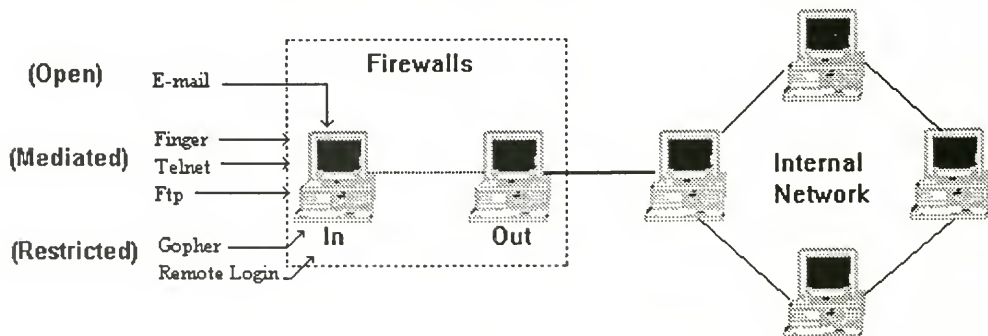


Fig. 2: Firewall Set-up

mailers can be very insecure. Users also need to log into various public archive sites to receive files. One solution to provide this type of functionality is outlined in Fig. 2, where two dedicated computers or gateways are used - one connected to the local network and the other to the Internet :

The external computer examines all incoming traffic, forwarding only "safe" packets to the internal machine. An attacker , thus, could only break into machines on the local network by compromising the internal gateway. The internal gateway only accepts messages from the external computer, so that, if unauthorised packets do reach it they will not be able to pass.

Conventional password techniques, when used with firewalls, reduce the efficiency of the overall security provided. Hackers gained access to Panix, a public-access Internet site in New York (Oct. 1993), and installed "packet sniffers". These programs watched data going by and recorded user names and passwords as people logged on to hundreds of other computer systems. Connections, therefore, within a firewall require, for ultimate security, a different kind of authentication mechanism that cannot be recorded by sniffers, e.g., "one-time password" or "challenge-response password".

Conclusions

Even if newcomers to the Internet try to secure their systems it's not always easy to find the information they need. Hardware and software vendors are usually loathe to discuss their security problems. CERT [23] generally issue advisories only after manufacturers have developed a definitive fix - usually weeks or months later. Spafford [10] points out that people don't know the risks - between half and three quarters of the security holes currently known to hackers have yet to be openly acknowledged. People only know the benefits. Many of these benefits come from programs such as Gopher, Netscape, World Wide Web or Mosaic, which allow simple menu-driven navigation of the Internet. These tools are drawing thousands of people to the net. Yet, the rapid evolution of these tools have bypassed steps that could lead to security breaches. The popular Gopher problem, for instance, according to CERT advisories, has security problems that make it possible to access not only public files but private ones as well. Again, Gopher is only insecure if it is misconfigured. Whilst Gopher servers can be confined so that they have access only to public information, by default they have a free rein.

The Internet attracts more recruits every day hoping to reap such rewards and benefits as connections with other people and organisations, file access and information interchange. Yet, Internet connection can also prove to be the source of very real risks and dangers to your workstation or network. While this article, hopefully, raises the reader's awareness of the importance of security in a network situation, one shouldn't get paranoid. It's worth remembering that security, in most cases, is an elusive goal. Don't forget as well that the nature of Unix and the Internet helped to defeat the Internet worm as well as spread it.

The sensible approach is to secure your system according to its needs, keeping danger at a manageable level. In other words, don't stop travelling but do wear a seat belt.

References

- [1] Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada.
- [2] "Data on Ethnic Conflict and Conflict Resolution : The Work of INCORE", Journal of Ethno- Development, Forthcoming Publication
- [3] Metadata on Ethnic Conflict and Conflict Resolution - INCORE Metadatabase Project, Iassist Quarterly, Forthcoming Publication
- [4] Kay, R. : "Distributed and Secure", BYTE, pp.165-178, (June 1994)
- [5] MIDS, "MIDS Press Release : New Data on the Size of the Internet and the Matrix", Matrix News, Vol. 5, No. 1, 1.5. Matrix Information and Directory Services, Inc., Austin, (January 1995), Available from mids@tic.com
- [6] Lottor, DNS ZONE SURVEY (RFC 1296) through January 1995
- [7] McLellan, V. : "NASA Hackers : There's More to the Story", Digital Review, p. 80. (Nov. 1987)
- [8] Spafford, E.H. : "The Internet Worm Program : An Analysis", Computer Communications Review, 19, No. 1, ACM SIGCOM, (January 1989)

- [9] Seeley, D. : "A Tour of the Worm", Proceedings of 1989 Winter Usenix Conference, Usenix Association, San Diego, California, (February 1989)
- [10] Eichin, M.W., Rochlis, J.A. : "With Microscope and Tweezers : An Analysis of the Internet Virus of November 1988", Proceedings of the Symposium on Research in Security and Privacy, IEEE-CS, Oakland, California, (May 1989)
- [11] Garfinkel, S., Spafford, S. : "Practical Unix Security", O'Reilly & associates (1992)
- [12] Todino, G., Strang, J., Peck, J. : "Learning the Unix Operating System", O'Reilly & Associates (1991)
- [13] Morris, R., Thomson, K. : "Password Security : A Case History", Communications of the ACM, 22(11), pp. 594-597, (November 1979). Reprinted in Unix System Manager's Manual, 4.3 Berkeley Software Distribution, University of California, Berkeley (1986)
- [14] Grampp, F.T., Morris, R. : "Unix Operating System Security", AT&T Bell Laboratories Technical Journal, 63 (8), pp. 1649-1672, (October 1984)
- [15] Schweitzer, J.A. : "Managing Information Security. Administrative, Electronic and Legal Measures to Protect Business Information", 2nd. ed. Massachusetts, Butterworth Publishers (1990)
- [16] Farmer, D., Venema, W. : "Improving the Security Of Your Site By Breaking Into It", Available by ftp from win.tue.nl as /pub/security/admin-guide-to-cracking.Z
- [17] Schuba, C. : "Addressing Weaknesses in the Domain Name System Protocol", Purdue University (August 1993)
- [18] Reinhard, R.B. : "An Architectural Overview of Unix Network Security (Specifically oriented towards Internet Connectivity)", v.4 (Feb. 1993), Available from gopher.near.net in /security/papers
- [19] Available via anonymous FTP from cert.org in ~/pub/tools/cops.
- [20] Available via anonymous FTP from athena-dist.mit.edu in ~/pub/kerberos
- [21] Bellovin, : "Limitations of the Kerberos Authentication System", Available via anonymous ftp from research.att.com
- [22] Ranum, M. : "Thinking About Firewalls", Proceedings of the 2nd. International Conference on Systems and Network Security and Management, Available via anonymous ftp from ftp.tis.com as /pub/firewalls/firewall.ps.Z
- [23] The Computer Emergency Response Team (CERT) advisories are available by ftp from cert.org

FOOTNOTES * a worm is an independently operating program that actively propagates by spreading copies of itself throughout a network.

@ In Usenet parlance a "*hacker*" is a person possessing a great deal of knowledge and expertise, and exercises this with great finesse, whereas a "*cracker*" is a person who persistently breaks into other people's computer systems, for a variety of reasons. For further information refer to Steele, G.L., Woods, D.R., Finkel, R.A., Crispin, M.R., Stallman, R.M., Goodfellow, G.S. : "The Hacker's Dictionary", New York: Harper and Row, 1988.

a *shell script* is a Unix program containing a series of commands that perform system functions.

& a *block of undesired code* (intentionally hidden within a desirable block of code) which does things that the user does not intend, e.g., a program that simulates a computer's logon procedure, but, rather than logging the user on, it simply records and steals the user's id and password.

\$ a feature built into programs by the designers allowing them special privileges which are denied to the normal users of the program, e.g., a *back door* in a logon program would enable the designer to log onto a system without an authorised account.

Locating and Accessing Data and Information on the Internet: Methods and Organizational Impacts

by Christopher Davis¹
CIESIN

"But it's really an information ocean, not a highway. If you think of it as an ocean, then you have to consider the kind of tools that are used, who builds the boats, who designs them, and whether you're surfing or diving. If you have a message in the bottle, how do you get the bottle to the people who need it?" Peter Gabriel, New York Times, July 13, 1994

The Internet provides a global infrastructure for data and information publishing that has the potential to revolutionize how data and information are accessed and used. While a variety of methods exist for taking advantage of the capabilities of the Internet for this purpose, two problems with data and information access have been exacerbated by an explosion of tools and resource servers. First, different tools and approaches each have individual advantages, leading to the use of different methods at different locations. Second, resources are distributed across many locations and may often be difficult to locate. CIESIN's information system approach is designed to solve both of these problems by providing a method for locating and integrating heterogeneous, distributed data and information resource servers. The implement of technologies such as those utilized by CIESIN could potentially transform the organization of how data and information are provided and provide users and providers alike with new services and capabilities.

The technical core of the Internet is a set of protocols and standards for computer to computer communication. Initially, Internet technology supported three primary high level services: access to remote systems (telnet), exchange of files (ftp), and electronic mail (smtp). Over time more sophisticated standards and protocols have evolved that offer other services as well. This technology allows the creation of client-servers systems capable of greatly enhancing the dissemination of data and information resources. Data and information providers have a variety of types of resource servers to select from, each with its own functionality (see Table 1).

Server	Functionality
World Wide Web (WWW) Hypermedia Documentbase	Browsing and retrieval of formatted text, images, sounds, movies combined in hyperlinked documents distributed across servers
Gopher Documentbase	Browsing and retrieval of text and/or graphics files in hierarchical lists distributed across servers
WAIS Index	Full text searching and retrieval of textual documents and/or graphical files
Database	Querying of and access to data structured by fields and records
Application	Data processing and analysis of data sets utilizing the capabilities of applications such as ARC/INFO and SAS
FTP Archive	Retrieval of text and/or binary files from a hierarchical list
Newsgroups/Mailing List	Multi-party, free-form, asynchronous textual discussion

Table 1. Internet Server Functionality

From the user perspective, each of these servers can be accessed

at least by a client application that matches the server. Some clients, though, provide access to multiple server types (see Table 2).

	WWW	Gopher	WAIS	Database & ² Application	FTP Archive	Newsgroup ³
WWW Browser	X	X	X	X ⁴	X	X
Gopher Client		X	X		X	
WAIS Client			X			
Database Front End				X ⁵		
FTP Client					X	
News Reader						X

Table 2. Internet Client Functionality

Because of the wide range of functionality offered by WWW browsers and servers, these systems are more widely used than any other approach for both access and distribution.

While WWW browsers offer access to a range of servers, the WWW architecture severely limits design flexibility in information systems. Current WWW standards such as HTML (hypertext mark-up language) and HTTP (hypertext transfer protocol) limit user interface design to what can be accomplished using a basic form interface. This precludes the use of features such as menu bars and dialog boxes and other dynamic windows. This problem is compounded by the fact that HTTP is a stateless protocol. The WWW browser requests a document; the document is provided by the server, and the connection closes. The server has no way of knowing or tracking what document the user requested once the request is filled, thus global settings and variables are difficult to maintain from one document to another. Also, the WWW browser is a display tool only. The browser has no capabilities for data processing, so all data processing must be performed at the server. This is potentially problematic in two instances. First, the server might become overloaded processing multiple jobs, which could be more easily handled by the client. Second, when an image such as a chart is created, the data sent to the WWW browser is a graphic file which will be a much larger file than the actual data used to generate the image. If the client uses data from the server to create an image, the amount of network traffic would be significantly reduced, leading to improved performance. A final problem is that while WWW browsers support some types of servers without modifications to the server, database and WAIS servers require development work on the server to allow access. Multiple options are available and documented for WAIS servers, and several examples exist of providing access to Oracle and other relational databases, but access to database and custom servers may involve significant server modification and development work. In some instances, it may not be feasible to develop an interface from WWW because the server requires a stated protocol. Unfortunately, WWW browsers alone are not a universal solution for Internet server access.

However, the advantages of utilizing WWW browsers and servers in information system design should not be minimized. From the development perspective, WWW browsers exist for all major operating systems, and the development of clients can be anticipated to continue as new operating systems emerge. This removes a major cost of information system development and support. From the user perspective, WWW browsers provide access to a range of services and are thus more likely to be installed and used on a regular basis than custom clients for a particular information system. This creates a major incentive for resource providers to utilize WWW servers since they are immediately available to the existing and massive install base of WWW browser users.

The development of applications such as WWW have facilitated the distribution of data and information resources over the Internet. While this has led to an explosion of available resources, a specific resource may be often difficult to locate. A frequent quote overheard on the Internet is: "Everything you need to know is on the Internet. You just can't find it."⁶ By design, the Internet is a cooperative, unmanaged venture. This is one reason why the Internet has been so successful, but it is also the source of the grand challenge of locating resources.

Several efforts have been undertaken to chart the Internet or at least provide mechanisms for facilitating the location of data and information resources. One listing of these efforts includes eighteen different searchable catalogs of Internet resources.⁷ Some efforts focus on cataloging specific types of resources. For example, the Council of European Social Science Data Archives (CESSDA) is developing an interface to the distributed collections of European and other social science data archives. At present, this involves a global map of resources with pointers to individual archive resources.⁸ The U.S. federal government has created the Government Information Locator Service (GILS)⁹ to facilitate the cataloging government resources using a standard system. The G-7 countries have agreed to prototype the GILS standards for non-U.S. resources as well, but the GILS project is still very much in the early phases of development. Another effort is being undertaken by the Consortium for International Earth Science Information Network (CIESIN). CIESIN has developed the CIESIN Gateway as a system for searching distributed metadata collections and providing access to heterogeneous resource servers.¹⁰ While these projects are still in development or early phases of implementation, the existing results show that the technology exists for solving the problem of locating resources on the Internet.

One of the approaches for locating data and information resources that has been implemented is the CIESIN Gateway. The CIESIN Gateway was initially developed to meet the requirements of CIESIN's mission as the SEDAC (Socioeconomic Data and Applications Center) in NASA's Earth Observing System Data and Information System (EOSDIS). One of SEDAC's charges is to serve as a two-way gateway between social science and physical science researchers studying global change. The CIESIN Gateway fulfills this function by providing a single interface that allows searching of multiple data archives. The CIESIN Gateway includes a single interface that allows searching of the EOSDIS IMS, NASA Global Change Master Directory, and related directories of data, and also allows searching of key social science and other related metadata collections. Organizationally, CIESIN accomplishes this task through its Information Cooperative program. The Information Cooperative provides an institutional umbrella for linking together data centers worldwide. The CIESIN Gateway provides the technical implementation for the Information Cooperative by allowing searching of the directories of each of those data centers.

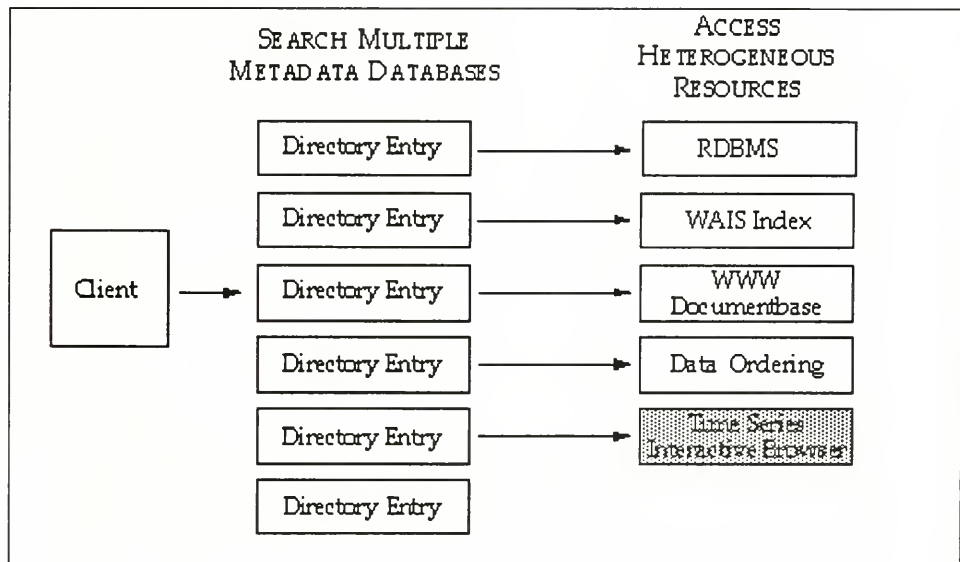


Figure 1. CIESIN Gateway Conceptual Schematic

The philosophy of the Information Cooperative is to encourage each partner data center to maintain their own metadata directories and data archives. This requires that the CIESIN Gateway be a distributed system, capable of searching multiple systems in parallel. Also, since each data center is likely to have different existing information systems, the CIESIN Gateway must support access to a variety of commercial databases and other Internet servers. In addition to searching and displaying high level metadata, the CIESIN Gateway also provides the capability of accessing on-line resources identified by the metadata. In some instances this is done within the CIESIN Gateway client, but in others it is accomplished by spawning an external application such as a WWW browser. A conceptual view of the capabilities provided by the CIESIN Gateway are described in Figure 1.

Thus, the CIESIN Gateway combines the capabilities of a search system for locating resources with capabilities for accessing those resources regardless of server type.

Based on the lessons learned from the development of the original CIESIN Gateway system, CIESIN, in collaboration with Brooklyn's Polytechnic University, is working on a new project, Raven. Raven subsumes the CIESIN Gateway functionality within a larger framework of access services. Raven presents the user with a list of services, such as CIESIN Gateway, WWW browser, and interfaces to custom applications and databases (see Figure 2).

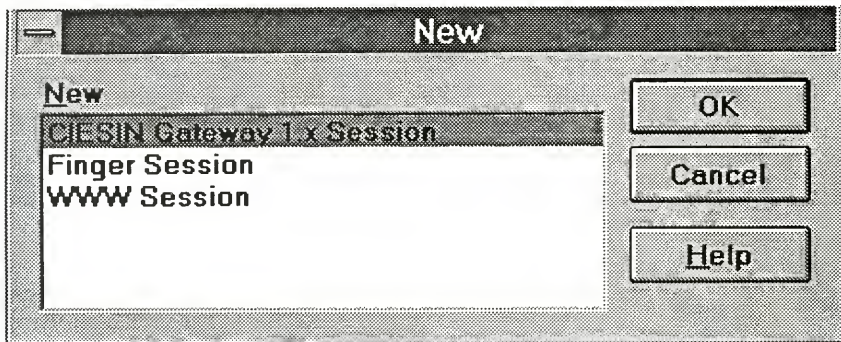


Figure 2. First Screen from Raven Prototype

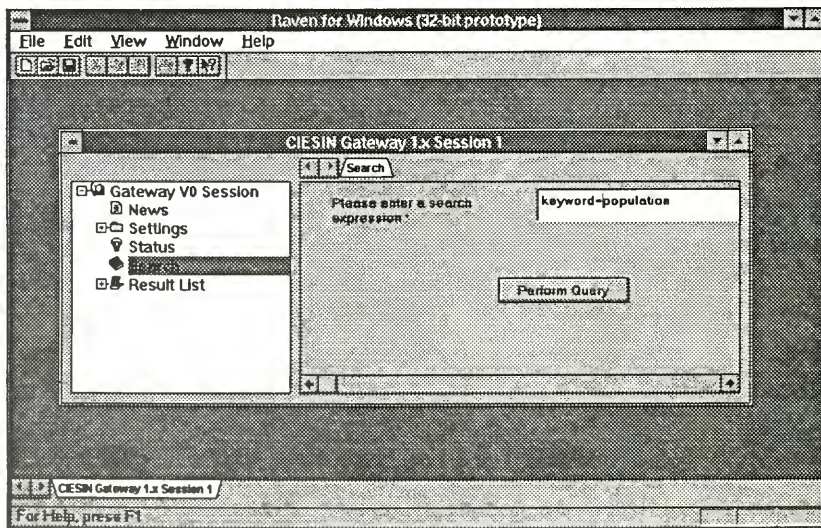


Figure 3. Search Screen from Raven Prototype

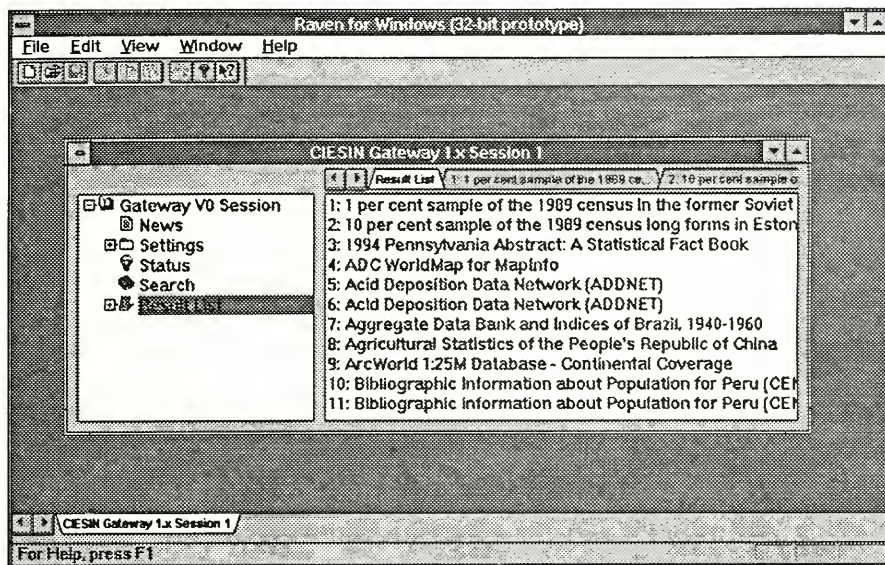


Figure 4. Search Results Screen from Raven Prototype

This approach provides more functionality within one client, and it also enhances the interoperability between separate services. The Raven design and development is based on object oriented design techniques that allow new services to be easily built using common components from other services. At the most basic level this includes the networking component of the services, but it also includes capabilities for viewing tables and entering queries (see figures 3-4).

Raven also uses cross platform development tools to facilitate the development of versions for multiple operating systems. The goal of the Raven project is to create a client that provides a single interface to a vast array of resources and systems and provides information system developers with a set of tools for easily creating a user interface appropriate for a particular system.

Technology such as WWW and the CIESIN Gateway provide the technical capabilities that allow organizations to take advantage of the infrastructure of the Internet to distribute data and information resources. These capabilities can have several impacts on the organizations that utilize these technologies. First, the Internet infrastructure allows for the development of new products and services that allow the creation of digital libraries. Digital libraries include network accessible collections of data and information resources. Because these resources are available over the Internet, the physical location of the user and the library itself becomes irrelevant. Also, digital, on-line storage of resources allows enhanced tools for searching, accessing, and analyzing resources. Thus, data and information resource providers can utilize Internet technologies to expand their user base and the services they provide.

A more subtle but significant organizational impact of these technologies affects the very organization of data centers. An example of this is the structure of CIESIN's Information Cooperative. To the user, the Information Cooperative appears as a single archive, but in fact, it is a collection of archives linked through the Internet and the CIESIN Gateway. This organizational approach is a form of adhocracy, a term first coined by the futurist Alvin Toffler in his book *Future Shock*¹¹. An adhocracy is based on small, specialized organizations that use information networks to coordinate their activities, and thus act like a larger organization. This approach can be more efficient and responsive than traditional hierarchical structures. This change is not limited to data and information resource providers. Similar changes are being experienced in a variety of industries, both public and private. As the global economy becomes increasingly information based and dependent, this pattern will increase in frequency. As with other types of organizations, two primary organizational roles emerge: providers and brokers¹². Providers develop and disseminate resources and provide support on the use and understanding of those resources. Brokers work with providers to develop catalogs of resources across providers and to

develop information systems to offer a common access system for users to the resources of multiple providers (see Figure 5).

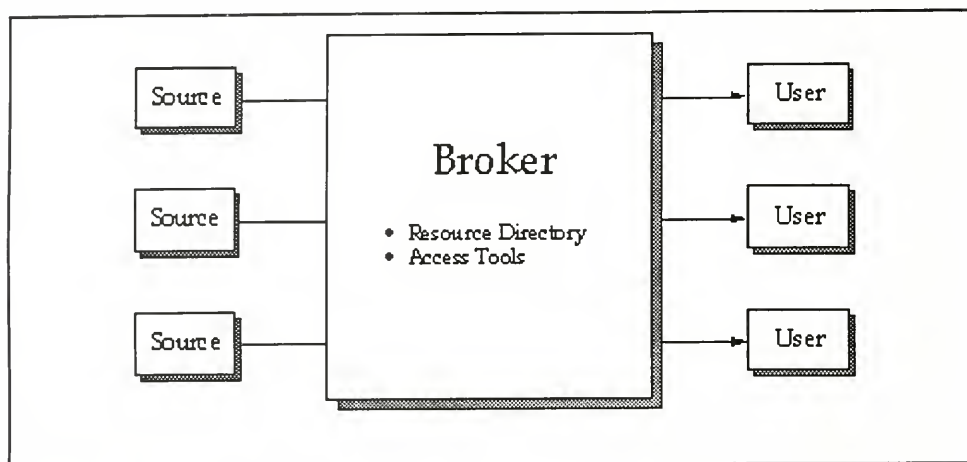


Figure 5. Relationship between Providers, Brokers, and Users

CIESIN, through its Information Cooperative program, represents one early effort at this approach. The CESSDA effort is another example. As the technical infrastructure for these types of organizational relationships becomes more widely implemented, other similar efforts will also emerge.

The Internet and related technologies have already had a significant impact on the practices of distributing and accessing data and information resources. Consider this closing thought from Christopher Locke (1994), "Unlike any previous medium, the Net's speed and reach seem to enable reaction to events that have not yet taken place. But this is an illusion. We are not seeing into the future, but more deeply into the present."¹³ This capability will force data and information resource providers to continue to take advantage of new methods both of technology and organization to enhance and improve speedy and effective access and use of data and information.

1 Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada.

2 The methods of user access for Database and Application servers are functionally equivalent.

3 Mailing lists are accessed through electronic mail and are functionally separate from any of the other client/server access / distribution methods.

4 The database interface is limited by the forms capability of the HTML standard and the existence of server side scripts for translating form input into a format understood by the database server and converting the results from the database and converting it to HTML.

5 A database front-end might be a separate client that runs on a users machine, or it might be a service that is accessed via telnet over the Internet. Generally, the front end is a custom interface to an off-the-shelf or custom server, usable only with a particular information system.

6 From David Lubar's "It's Not a Bug, It's a Feature": Computer Wit and Wisdom, Addison-Wesley Publishing, Reading, MA, 1995 who cites the source as "Anonymous, but common knowledge to anyone who's been there."

7 <http://cuiwww.unige.ch/meta-index.html>

8 <http://www.uib.no/nsd/diverse/utenland.hum>

9 <http://www.usgs.gov/gils>

10 <http://www.ciesin.org/gateway/gw-home.html>

11 Random House, New York, 1970.

12 An excellent summary discussion on this is "Electronic Markets and Electronic Hierarchies," by Thomas W. Malone, Joanne Yates, and Robert I. Benjamin in *Computer-Supported Cooperative Work: A Book of Readings*, edited by Irene Grief, Morgan Kaufman Publishers, San Mateo, CA, 1988. Additional references are available from the author of this paper.

13 From David Lubar's "It's Not a Bug, It's a Feature": *Computer Wit and Wisdom*, Addison-Wesley Publishing, Reading, MA, 1995.

SOSIG - a move towards subject-based services.

by Nicky Ferguson¹

*ESRC Visiting Fellow In Networked Information
Social Science Information Gateway Project,
University of Bristol, England.*

HOLDING HANDS OR OPENING GATEWAYS?

The UK was, and is, very fortunate in that an early commitment was made to providing a nationwide integrated academic computer network - JANET. This made it easier for successful national initiatives to be devised and implemented. However JANET was based on X25 protocols, not the Internet protocols known as "IP". This hindered international integration. It became clear that arguments over the relative merits of protocols were irrelevant because the Internet was going to be the de facto standard for international networking. So the UK started a rapid transition to driving on the same side of the road, in networking terms, as the rest of the world. Fortunately the existence of JANET means that, for academic institutions, this transition will be completed in a relatively short time.

In June 1992, some time before this transition was decided upon, but when it was already looking inevitable to many of us, I was appointed by the UK Economic and Social Research Council with a brief to support UK Social Scientists in the use of computer networked information. This was, and is, a very broad brief; and since there was not a precedent for a job of this kind I was to some extent improvising, making the rules up as I went along.

The infrastructure for networked communication had been built by the technicians. So, initially, it was the technicians who tended to use it. It wasn't until the infrastructure was there, and the user base expanded significantly away from the original designers and builders, that the way people were going to use it "in real life" began to emerge. It no longer seems surprising to us that a medium designed for the rapid exchange of large data files for "serious" work is now most popularly used for exchanging short messages. It may well be that an infrastructure intended for file transfer of software and complex datasets will be used mainly for magazine publishing and the promotion and delivery of new service industries. As it has been with infrastructure so with information. Whilst the providers and users were the same small band it was not clear what the problems or the potential were in this area. Moreover, the provision of networked information has largely been the realm of the technical specialist, not the information specialist. There are clues here to understanding the subsequent rather uneven development in this area. It has become easier to provide information, and much development has gone into the *interface* for users, with tools such as Mosaic and Netscape, but the information processing side has lagged behind.

I entered this arena at a time when there was clearly sprouting enthusiasm amongst social scientists who had previously regarded computers with fear and thought of email as just another way of increasing their workload. This enthusiasm was fed by global consciousness-raising in many fora and by my own small efforts. But often this enthusiasm did not progress from my visit - it did not translate into "real work". It was fine to go step by step through the maze with hand-holding documentation and a supportive guide but not so easy to navigate to unknown territory after a few weeks had elapsed. Even for the brave there were more obstacles - the origins of the infrastructure mentioned above meant that relevant social science information was scattered and sparse, often seeming to occur incidentally. My workshops and demonstrations were offering a glimpse of the possibilities rather than handing out a tool which could immediately increase the efficiency and productivity of researchers. It was all very well to look at all this fancy stuff, and the feedback from my sessions was always very good, but people still regarded this "Internet stuff" as a plaything.

As information provision slowly became easier, some academics were of course delighted to rediscover the joys of the second hand bookshop. Spending an hour browsing the networks might, or might not, uncover the odd jewel amongst the dust and chaos. Once found, the jewel could be copied or printed out and squirrelled away with the other printouts and photocopies. But only the most organised users made a note of their path as they went, so after a few days directing a colleague to rediscover the jewel might be impossible. Even now with the facilities of browser programs one person's hot list is another person's cold shower. I'm not too interested in the schedule for evening classes in a college in the mid-west of the USA - but it might be very useful to the right user. Making some personal details about oneself available over the net might perform an important function - to personalise and humanise discourse in a collaborative project where the participants have never met, for example. But I do not want to keep tripping over these details when I'm looking for something else. So system administrators, responding to complaints, hit upon the extraordinary idea of organising information according to subject headings. It took a little while for everyone to realise that librarians have been doing something similar for years, and by that time a host of idiosyncratic infant subject classification schemes were sprouting. In setting up the Social Science Information Gateway, we resolved to attempt at least to share the underlying classification system with

other UK national service providers.

THE RETURN OF THE CATALOGUER

With the advent of client or browser software giving users a graphical interface to networked information, the possibilities for junk or vanity publishing seemed to expand dramatically, the idea of making pictures, text and sounds available across the world -do-it-yourself multi-media publishing - was irresistible. Combine this with the relative ease of creating *HTML* - hypertext mark-up language, the building block of the World Wide Web, and you have an explosive combination. While the development of publishing was bounding ahead, the users of information were not so well provided for. Browsing was more exciting - instead of showing users meteorological data in tables, I could now bring up on their screens satellite photographs in glowing colour. But even with subject categories and fancy graphics, all we have really done is to give the second hand bookshop a facelift. You may know which shelf to look on, if you're lucky some of the books may have glossy covers, but the essential problem of locating relevant and useful texts remains.

One way we have tried to deal with this at SOSIG is by providing a searchable catalogue of information about each of the over 500 resource centres at which we point. This is quite different from the various so-called robots or automated search mechanisms which rely on highly resource intensive scouring of the networks and fairly crude automated examination of the resources themselves. We rely on human intervention to describe, classify and organise social science resource centres. In this way we also introduce an element of quality control. For each resource centre which appears anywhere on the subject menus, a form or template has been filled out - this contains a description and keywords as well as appropriate technical information such as the URL (network address) and the UDC (classification) number assigned to that resource. The user can then search through this information using an on-screen form. A dynamic list of hits will then be returned listing appropriate resource centres, describing them and pointing directly to them. Various options are provided and others (including Boolean search options) will be added in the near future.

ROADS TO THE FUTURE

The ideal for such services is that they should be distributed - so that centres of expertise are responsible for relevant subject areas. To answer the obvious question that this raises about our own activities, it is probably neither feasible nor desirable in the long term for us to attempt to take responsibility for describing and organising all the social science resource centres in the world, it is surely better for centres of excellence within the different social sciences to take responsibility for their own areas and for us to co-ordinate these efforts, but as a medium term solution the current SOSIG is certainly preferable to a totally centralised

model. Aiming for a distributed model, however, creates its own problems. It demands the ability to search across different servers which in turn implies that the resource descriptions will be in (preferably an internationally accepted) standard form. When the catalogue databases become large, as they undoubtedly will, manipulating the descriptions and templates will also become a problem if we rely on the relatively unsophisticated tools we use at present. In addition this system of describing and searching for networked resources should not be idiosyncratic - it should be adaptable and aim for future integration with other resources such as OPACs and citation indices.

For these reasons, in collaboration with UKOLN, the UK Office for Library and Information Networking at the University of Bath, and Loughborough University of Technology, we have recently been funded to develop a system for allowing linked and geographically distributed resource discovery services to be set up.

ROADS -Resource Organisation and Discovery in Subject-based services- will allow users to search across different subject-based servers and will develop searching mechanisms based on emerging Internet standards such as Whois++. It will also investigate integration with other standards such as Z39.50 and Marc (in its various incarnations). As well as expanding the knowledge base and the capabilities of services such as SOSIG, *ROADS* will provide a packaged solution for information providers who wish to set up a subject-based service. We also hope to encourage centralised national service providers to focus their effort on the (initially many) subject areas not covered by these distributed services, so that a good coverage can be achieved in a relatively short time. Thus *ROADS* will help to achieve the goal of a scaleable system for resource discovery, cataloguing, description, organisation and quality control.

We have no illusions that *ROADS* will be a so-called killer application for networked information - there will not be such an application, rather a number of different approaches will emerge and possibly merge. Moreover sometimes a user will not find the obscure object of desire; or perhaps wishes to *comprehensively* survey networked resources on a topic without necessarily having regard to quality or currency; or to search across different languages and character sets. For these reasons, the *ROADS* partners intend to collaborate with European partners, not only to develop *ROADS* further, but also to develop complementary systems, including a comprehensive automated indexing system for European World Wide Web servers. Thus, if the ergonomic nut crackers fail to break open the shell and reveal the kernel, we will provide the back-up of a well-designed hammer.

This European collaborative proposal, codenamed *DESIRE*, has recently been shortlisted for funding by the relevant European funding agency.

We hope that all three of these initiatives will promote the design and building of Subject-based Information Gateways (SBIG's), the implementation of which will result in a distributed resource discovery service based on rich descriptions and a quality controlled approach organised around subject centres of excellence. These efforts will be complemented by a comprehensive approach to European WWW index design, the implementation of which will result in a European discovery service based on automated indexing and an automated harvesting technology.

REFERENCES AND FURTHER INFORMATION

1 Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada.

BUBL - The Bulletin Board for Libraries -

<URL: <http://www.bubl.bath.ac.uk/BUBL/>>

FIGIT - The Follett Implementation Group on Information Technology -

<URL: <http://lamin.bath.ac.uk/FIGIT/figit-2-94.html>>

Guardian (1995). *Online* March 30, 1995

NISS - National Information Services and Systems -

<URL: <http://www.niss.ac.uk/>>

ROADS - Resource Organisation and Discovery on Subject-based Services -

<URL: <http://ukoln.bath.ac.uk/ukoln/roads/roads.html>>

SOSIG - Social Science Information Gateway -

<URL: <http://www.sosig.ac.uk/>>

UKOLN - The UK Office for Library and Information Networking -

<URL: <http://www.ukoln.bath.ac.uk/ukoln/>>

Further reading may be found at:

<URL: <http://ukoln.bath.ac.uk/ukoln/roads/related.htm>>

WWW: What do Researchers Want?

Summary of 44 Responses to an E-mail Survey February-March, 1995

by Jim Henderson¹
Maine State Archivist

Introduction

During February, 1995, I distributed a brief survey, "WWW: What do Researchers Want?", to approximately thirty history and other listserves. Others were approached, but not all allow non-subscribers to use their lists. A single "reminder" e-mailing was sent in March. Each distribution generated just over 20 responses, for a total of 46. All responses were received electronically.

In brief, researchers want clear guides to collections, supplementary information about the institution and its mission, and access information: rules for copying; mail, phone, e-mail information. They are far less interested in "cute" sample images (the old map or photo) or sample text of selected collections. Parochial items such as organizational structure or exhibits and upcoming events are clearly the lowest priorities among those listed.

Researchers most highly value "subject oriented keywords pointing to related collections," and "detailed descriptions" and "finding aids" for major collections. Next, they want to know the ways and means of access: rules about the cost and availability of copies, both traditional (mail, phone) and e-mail contact information.

After the basics, and to get a view of the institution's possibilities, researchers want 1) listings of collections by genre, 2) lists of guides, pamphlets, and other publications, supported by 3) reference room hours and procedures, and 4) a general description of the institution's holdings and mission. Following closely are interactive needs: the ability to leave messages for the staff and to find out "What's new?"

While given "some importance," image databases of photos and maps were deemed slightly less useful than the proposed textual databases, which also were not highly sought after. Selected sample items, by both typical content and format, were viewed unfavorably by one-third of the respondents.

Internal and local items characterized by "organizational structure" and "upcoming events" received rather negative reviews. Current research listserve members find little interest in genealogical holdings, but this may say more about the respondents and the current availability of technology than about the potential broad interest in this information.

The Detailed Responses: an Analysis

Respondents were asked to rank the proposed features as *Very Important* (V), *Important* (I), *Some Value* (S), *Not Important* (N), or *Forget it* (F). The first three columns at the right below display the percentage of responses to the two lowest ratings (FN), the middle (S), and the highest ratings (IV). (Rounded percents may not add to 100.)

The next column reflects the average of all responses, with *Very Important* coded as 4; *Important*, 3; *Some Value*, 2; *Not Important*, 1; and *Forget it* as 0. The ranking of average scores sometimes differs from the order of the highest ratings because of 1) the varying portion of respondents choosing "*Very Important*," etc. as their selections, and 2) the fact that a few respondents did not respond to all items.

The final column notes the standard deviation from the mean of all responses. The lower the number, the greater cohesion among respondents, with a tendency to cluster about one of the five choices. Standard deviations ranged from .68 to 1.11.

Highest Rated Features

The most desirable features center around three themes: collections level descriptions, availability of copies, and contact information. While "subject oriented keywords" were most valuable, "detailed descriptions" and "finding aids" of major collections were highly regarded. Rules concerning the cost and availability of copies ranked second overall, while requirements to have both traditional (mail, phone) and e-mail contact information were equally valued as high priorities. All features in this group had average scores tightly clustered from 3.2 through 3.4. Except for the "mailing, location," item, they also had relatively low standard deviations. Basically, these features are highly rated, with over 80% endorsement as

Important or *Very Important*, and represent the relatively uniform opinion of researcher respondents.

Proposed Feature	Percent				
	FN	S	IV	Av	SD
9. Subject oriented keywords pointing to related collections	2	9	89	3.2	.78
4. Availability, cost, restrictions regarding copies of records	0	12	88	3.3	.68
8. Detailed description of major collections	2	11	86	3.4	.69
19. Finding aids for major collections	7	7	86	3.3	.79
2. E-mail addresses of site and key staff/departments	2	16	82	3.4	.75
1. Mailing address, location, telephone number of the site	2	16	82	3.3	.95

Table 1

Majority Supported Features

After a fairly clear break of 15 percent in the *Important/Very Important* ratings, the following appear to be “helpful, supplemental” features. These features all rank from 3.0 (*Important*) to 2.5 (*Important/Some Value*).

To get a view of the institution’s possibilities, researchers want 1) listings of collections by genre, 2) lists of guides, pamphlets, and other publications, supported by 3) reference room hours and procedures, and 4) a general description of the institution’s holdings and mission. At 3.0 and 2.9, these essentially rate *Important* on average.

Following closely are interactive needs: the ability to leave messages for the staff and to find out “What’s new?”

Rather lower in this group’s ranking (and close in content and rank to the first two features in the next section) are requests for textual databases describing photographic and cartographic holdings publications articulate a cluster of desirable features.

Proposed Feature	Percent				
	FN	S	IV	Av	SD
21. List of collections by genre: text, map, photo, video, audio	9	23	67	3.0	.93
6. List of guides, pamphlets and other publications	0	34	66	3.0	.85
14. Ability to leave message for archives staff	5	30	65	2.8	.81
20. What New? (acquisitions, finding aids)	9	29	64	2.8	.95
5. Reference Room hours, procedures, rules	16	18	66	2.9	1.02
7. General description of holdings and mission - 1-2 pages	5	36	59	2.9	.99
15. Database of textual description of photographs	20	25	55	2.5	.96
16. Database of textual description of maps	20	25	55	2.5	.96

Table 2

Lowest Rated Features

The final six features had no majority expression of combined *Important/Very Important* responses. All rank below 2.5 on average and cluster around the *Some Value* rating of 2.0..

Image databases of photos and maps were deemed slightly less useful than the proposed textual databases in the previous section, though some comments indicated a “nice, but utopian” view of the proposition. Selected sample items, by both typical content and format, were viewed unfavorably by a third of the respondents.

Internal and local items characterized by "organizational structure" and "upcoming events" received rather negative reviews. Interestingly, while the overall ranking of these two items is similar, the standard deviation reveals virtual consensus (SD=.68) on the limited value of "exhibits, upcoming events," but a wide disparity of views (SD=1.11) on the value of organizational structure information.

Current research listserve members find little interest in genealogical holdings, but this may say more about the respondents and the current availability of technology than about the potential broad interest in this information.

Proposed Feature	Percent				
	FN	S	IV	Av	SD
17. Image database of photographs	20	34	45	2.4	.87
18. Image database of maps	20	39	41	2.4	.87
3. Description of organizational structure.	32	30	39	2.2	1.11
10. Selected sample items, major collections, typical content	32	43	25	1.9	.91
12. Genealogy holdings summary	32	43	25	2.0	.87
11. Selected sample items, major collections, typical format	34	43	23	1.9	.91
13. Exhibits, upcoming events	27	64	9	1.8	.73

Table 3

Ranking by Average Rating

In yet another arrangement of responses, this time by average rating, similar conclusions may be drawn.

3.5	<i>Important to Very Important</i>	
3.4		Detailed descriptions / E-mail addresses
3.3		Info about copies/Major finding aids/Postal mail, location, phone
3.2		Subject oriented keyword searches
3.1		
3.0	<i>Important</i>	List of guides, publications / List of collections by genre
2.9		Reference room hours, rules / General holdings, mission
2.8		Leave messages for staff / What's new?
2.7		
2.6		
2.5	<i>Some Value to Important</i>	Databases of textual description: photographs and maps
2.4		Image databases of photographs and maps
2.3		
2.2		Description of organizational structure
2.1		
2.0	<i>Some Value</i>	Genealogy holdings: summary
1.9		Selected sample items indicating typical format and content
1.8		Exhibits, upcoming events
1.7		
1.6		
1.5	<i>Not Important to Some Value</i>	

Table 4

The Survey as Sent

Sorry for duplications. This survey has been posted to over 30 history lists. I have not posted to lists focusing on non-North American history. Feel free to post to other lists you think appropriate. Please respond by February 24th. Thanks.

WHAT DO RESEARCHERS WANT TO KNOW FROM ARCHIVAL SITES ESPECIALLY REGARDING WWW DESIGN?

The Maine State Archives is about to establish a WWW site and series of "pages." The last few months have seen an explosion in this area. While we have reviewed many of the new sites, we wonder "What do researchers want to know?" Here's your chance! I will post this survey's results on the Archives Listserv and anywhere else it may be helpful. Your responses may apply to GOPHER design as well. Keep in mind that not all wishes are granted - if everything is "very important" then

Please reply to ME - hendersn@saturn.caps.maine.edu -and NOT to the list on which this is posted!

Archives WWW Design Survey

How important are the following in an archival electronic information site?

(V)ery important (I)mportant (S)ome value (N)ot important (F)orget it!

1. Mailing address, location, telephone number of the site
2. E-mail addresses of site and key staff/departments
3. Description of organizational structure.
4. Availability, cost, restrictions regarding copies of records
5. Reference Room hours, procedures, rules
6. List of guides, pamphlets and other publications by the site, including ordering info
7. General description of holdings and mission - 1-2 pages
8. Detailed description of major collections: title, inclusive dates, summary, scope
9. Subject oriented keywords pointing to related collections
10. Selected sample items from major collections indicating typical content
11. Selected sample items from major collections showing typical format through displayed images
12. Genealogy holdings summary
13. Exhibits, upcoming events
14. Ability to leave message for archives staff
15. Database of textual description of photographs
16. Database of textual description of maps
17. Image database of photographs
18. Image database of maps
20. Finding aids for major collections
21. What New? (acquisitions, finding aids)
22. List of collections by genre: text, maps, photographs, video, audio

Places you have been (virtually), like, and why: _____

Additional comments, suggestions: _____

THANKS. PLEASE RESPOND BY FEBRUARY 24 (later MARCH) TO hendersn@saturn.caps.maine.edu

Jim Henderson, State Archivist

Cultural Building, Station # 84

Augusta, Maine 04333 (207) 287-5790

hendersn@saturn.caps.maine.edu

MAINE Archives BBS 207-287-5797

Comments from Respondents:

Places you have been (virtually), like, and why:

I use Congress and related gophers to collect data and the text of documents and bills as well as information of votes. Some servers provide campaign expenditure data which is useful to me. I sometimes connect with electronic collections (e.g. Gutenberg project). So the availability of raw and secondary data is useful to scholars like myself. A second useful area involves access to catalogues and directories. Often, it is sufficient to know that a document exists and what the source is without actually viewing the document. Similarly, directories of various types are useful.

Still exploring, but I liked the Oregon State University Site WPA Exhibit, the Cornell exhibit, and see incredible value to the Johns Hopkins Gopher site.

U. of Michigan MLink (gopher://mlink.hh.lib.umich.edu/) - lots of information and links to other sources, sensibly arranged and easily accessed.

UKANS- HISTORY research

Thomas, because you can go back and forth in your searches — nice menuing. The Star Trek site at Paramount has some nice moving around tools, too:

Places with a good, thought out design, graphics that transfer quickly or very few graphics at all.

Additional Comments - Planning:

I think it will be important to think through the various audiences you want to reach—not only now, but in the future. [Perhaps a review of the site's (organization's) mission statement would be helpful.] The site should be designed with that (those) audience(s) in mind, and people need to recognize that one structure may not meet the needs of all audiences or users. Having said that, let me add that collecting survey information from somewhat experienced web users is one good source of information. Another source might be focus groups with intended users. Also, if funds permit, you could draft a basic design, set up some pilot tests with various groups of intended users (librarians, teachers, students, others). Then watch what they do. See what they like and don't like. What's confusing and what seems to flow more naturally, etc. Then debrief through focus group interviews

The reason I think images have only limited importance is a) some people are still using text-based readers, b) images take a really long time to load, c) even at their best, you can't always tell if an image is what you want. This is especially true for maps. Also, if you are going to have messaging for staff, you need to have a really reliable system of responding. An explanation of the searching tools would be nice.

I would say that indexing and maybe full text of various state publications, periodicals, newsletters, and even local newspapers would be very helpful to many researchers. If full text is available I would say that indexing is unnecessary. Including local newspapers would be extremely helpful as I am sure you know that national newspapers tend to not cover Maine very much or very thoroughly. This would provide a gold mine for people who are doing research on Maine.

Search & preview capabilities should be maximized.

15-18. I am a bit confused on this. If you mean all photographs and maps, then that would be wonderful. Many researchers then would not even need to travel to a particular site. But that would also be a massive task for the people at a given site and could demand immense computer memory.

If, on the other hand, you mean only certain photographs and maps, would that be any different from what you mean in #10 and 11?

Or, do you mean detailed descriptions of the "collections" of photographs and maps in a given archival site, and if so, would that be much different from what you mean in #8?

Tourist information—hours, locations, costs, restrictions—are not necessary to me as a research scholar. I do need to know what you have. I may need to search your holdings by any word or combination of words in order to know whether I need to ask about hours and restrictions. Even more wonderful, would be the ability to scan texts to determine the value of documents.

Georeferenced Population Data

by Hendrik Meij and Robert Chen¹

Consortium for Earth Science Information Network

INTRODUCTION.

The primary mission of The Socioeconomic Data and Applications Center (SEDAC), at the Consortium for International Earth Science Information Network (CIESIN), is to develop new policy oriented applications and information products that synthesize earth science and socioeconomic data. The SEDAC policy applications development effort is the primary means by which the Earth Observing System Data and Information System (EOSDIS) program helps to ensure that the scientific investment embodied in NASA's Mission to Planet Earth (MTPE) program leads to tangible benefits. SEDAC's activities closely link with ongoing activities related to land use and trace gas emissions at other EOSDIS Distributed Active Archive Centers (DAACs) such as EROS Data Center (EDC) and Oak Ridge National Laboratory (ORNL). In addition, SEDAC's efforts also play a critical role in the arena of integrated assessment of global climate change. Its aim is making outputs of the U.S. Global Change Research Program (USGCRP) useful for policy and develop specific tools and mechanisms to enhance the use of integrated assessment models in the policy process.

Population dynamics and distribution have been consistently identified as key elements in understanding human interactions with the environment and in considering possible responses to environmental change. The National Research Council (NRC) has identified (1) population dynamics as one of five priority areas of research for the USGCRP. It also points out the key role of georeferenced social data in two other priority areas:

- 1) improving understanding of land use change, and
- 2) assessing impacts, vulnerability, and adaptation to global changes.

In addition, the report highlights the need to address the full range of energy policy options in relationship to greenhouse gas reduction. The 1992 report of the Human Dimensions of Global Environmental Change Programme (HDP) on Population Data and Global Environmental Change (2) emphasized the importance of georeferenced population data for global change research and applications and recommended development of data bases at several different levels of aggregation and resolution.

GEOREFERENCED POPULATION DATA.

Georeferenced population data provide a critical link between data on the natural environment and data on human behavior and welfare. Past and potential policy oriented uses of such data include natural resource management, famine early warning and vulnerability assessment, design and planning of sample surveys, damage assessment and associated disaster response, assessment of the impacts of environmental variation and change (e.g., coastal storms and sea level rise), public health and medical service applications, urban and regional planning, and estimation of pollution emissions and land use change.

Spatial and temporal data on population and environment can be useful in analyzing behavioral impacts. For example, energy analysts interested in emission reduction policies may want to consider population location and behavioral patterns in relationship to alternative energy sources, air and water resources, pollution control, work and recreational areas, and transportation infrastructure. Urban planners may want to understand the effects of decentralized versus centralized population growth on energy use and emissions and assess the effects of alternative population and land use policies on population distribution. National policy makers are likely to be interested in the causes and impacts of internal and international migration, especially as they relate to environmental degradation and change in specific regions. Educators may want to introduce such data sources early in the curriculum to provide a base for interaction with national and local processes.

DATA PRODUCTS.

SEDAC has been developing a number of different georeferenced population datasets:

- o A set of gridded population datasets for more than 120 countries originally developed by the Center for International Research (CIR) of the U.S. Bureau of the Census for the Department of Defense (DoD) and recently made available

through SEDAC. The data files contain both urban and rural population density data at a resolution of either 20 by 30 minutes or 5 by 7.5 minutes. These files may be accessed via anonymous File Transfer Protocol (FTP).

ftp ftp.ciesin.org user <ftp> password <email address> cd /pub/data/Global_Population_DB

o A suite of high resolution, one tenth of one degree, population data sets for the globe. Two gridded data sets are provided, one with smoothed population counts data, and one with smoothed population density data, both at the 1/10th of a degree resolution. These data are smoothed by applying a mathematical technique (pycnophylactic interpolation) for preserving areal data while redistributing such data on a sphere (3). These files may be accessed via anonymous FTP.

ftp ftp.ciesin.org user <ftp> password <email address> cd /pub/data/Global_Demog_Project

o A georeferenced population dataset for the conterminous U.S. at the square kilometer resolution level. The approach taken in the development of this product requires both input from the U.S. Bureau of the Census 1992 TIGER/Line and Summary Tape File 3A databases. The initial attempt will focus on the transformation of census blockgroup total persons counts and total housing units structures to the pixel based format of a square kilometer. Several disaggregation methods are applied including majority rule, pycnophylactic interpolation, proportional allocation, and geostatistical estimations (kriging). Prototype efforts may be found at:

ftp ftp.ciesin.org user <ftp> password <email address> cd /pub/census/usa/grid

Please email ciesin.info@ciesin.org for up to date information.

o SEDAC has identified available sub national administrative boundary files for most countries of the world and is developing an integrated product using a public domain version of Digital Chart of the World (DCW) at a scale of 1:1,000,000. Such boundary files are critical in making linkages between socioeconomic data (e.g., on population, land use, and energy production and consumption) and natural science data (e.g., on land cover, vegetation change, and pollutant levels).

ACCESS PRODUCTS.

A series of activities are underway to make data from the U.S. Bureau of the Census, and possibly other national census takers, more widely accessible and usable. This includes development of:

o An interactive tabulation generator publicly accessible over the Internet. This program enables rapid access to very large census datasets for user defined cross tabulations generated from microdata samples. Currently, the U.S. 1% Public Use Microdata Samples (PUMS) files for 1980 and 1990 are available. Results may be save and retrieved via Kermit, FTP or email.

Please email ciesin.info@ciesin.org for up to date information.

o An interactive extraction generator publicly accessible over the Internet. This program enables the user to define, interactively, custom extractions to be performed on very large census datasets such as the U.S. 5% Public Use Microdata Samples (PUMS) files for 1980 and 1990. A custom extraction file is generated with a custom codebook and data dictionary. Retrieval mechanism include only FTP.

Please email ciesin.info@ciesin.org for up to date information.

o Creation of an " Archive of Census Related Products" involving the generation of usable products from pre tabulated datasets, such as the U.S. Summary Tape File 3A. Boundary files (from Tiger/Line 1992). Demographic data and boundary files are accessible via anonymous FTP. The demographic data records are uniquely linked to the appropriate area entities in the boundary files. The current Archive contains 11,300 retrievable files spanning 4,000 MB.

ftp ftp.ciesin.org user <ftp> password <email address> cd /pub/census <indent> Informational 'readme' files are echoed to the screen to guide the user in navigating this Archive. WWW browsers, such as Mosaic and Netscape clients, can enter the Archive via:

[http://www.ciesin.org/datasets/us demog/us demog home.html](http://www.ciesin.org/datasets/us%20demog/us%20demog%20home.html)

Look for the “/pub/census” hypertext link.

- o The ability for visual display, mapping, of census data by coverages, for browsing/ inspection purposes before retrievals. <indent2> <http://www.ciesin.org:2222/map.html>

- o The generation of “Dataset Guides” for informational purposes. <indent2> [http://www.ciesin.org/datasets/us demog/ us demog home.html](http://www.ciesin.org/datasets/us%20demog/us%20demog%20home.html)

SUMMARY.

The development of these products, providing rapid access to large demographic databases, combined with visual presentation of the materials, might enable more focused response measures to catastrophic events and others. It is hoped that in the short term, users will express and identify databases of interest for merging into the prototype systems under development.

LITERATURE CITED.

1. Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada. Consortium for Earth Science Information Network 2250 Pierce Road University Center, MI 48710 (hmeij@ciesin.org, bob.chen@ciesin.org)
2. National research Council (NRC). 1994. Sciences Priorities for the Human Dimensions of Global Change. National Academy of Sciences Reports.
3. Human Dimensions of Global Environmental Climate Change Programme (HDP). 1992. Population Data and Global Environmental Change.
4. Tobler, W. 1979. Smooth pycnophylactic interpolation for geographical regions. Journal American Statistical Association, 74(367): 519-536.



INTERNATIONAL ASSOCIATION FOR
SOCIAL SCIENCE INFORMATION
SERVICE AND TECHNOLOGY

• • • •
ASSOCIATION INTERNATIONALE
POUR LES SERVICES ET
TECHNIQUES D'INFORMATION EN
SCIENCES SOCIALES

Membership form

The International Association for Social Science Information Services and Technology (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the IASSIST QUARTERLY. They also benefit from re-

duced fees for attendance at regional and international conferences sponsored by IASSIST.

Membership fees are:

Regular Membership. \$40.00 per calendar year.

Student Membership: \$20.00 per calendar year.

Institutional subscriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subscription:

\$70.00 per calendar year (includes one volume of the Quarterly)

I would like to become a member of
IASSIST. Please see my choice below:

- ☐ \$40 Regular Membership
☐ \$20 Student Membership
☐ \$70 Institutional Membership

My primary Interests are:

- ☐ Archive Services/Administration
☐ Data Processing
☐ Data Management
☐ Research Applications
☐ Other (specify) _____

Please make checks payable
to IASSIST and Mail to :
Mr. Marty Pawlocki
Treasurer, IASSIST
% 303 GSLIS Building,
Social Science Data
Archives, University of
California, 405 Hilgard
Avenue, Los Angeles, CA
90024-1484

Name / title

Institutional Affiliation

Mailing Address

City

Country / zip/ postal code / phone

SERIALS DEPARTMENT (SERLIBS82135346)
UNIV OF NORTH CAROLINA-CHAPEL HILL
CB #3935 DAVIS LIBRARY
CHAPEL HILL NC 27514-6590

JUN 23 1996
OH

101